Parametric modelling for single-channel blind dereverberation of speech from a moving speaker

Parametric modelling for single-channel blind dereverberation of speech from a moving speaker

Christine Evers and James R. Hopgood

Institute for Digital Communications, School of Engineering and Electronics

The University of Edinburgh, EH9 3JL Edinburgh, Scotland

{c.evers,james.hopgood}@ed.ac.uk

Abstract

Single-channel blind dereverberation for the enhancement of speech acquired in acoustic environments is essential in applications where microphone arrays prove impractical. In many scenarios, the source-sensor geometry is not varying rapidly, but in most applications the geometry is subject to change, for example when a user wishes to move around a room. This paper extends a previous model-based approach to blind dereverberation by representing the channel as a linear time-varying all-pole filter, in which the parameters of the filter are modelled as a linear combination of known basis functions with unknown weightings. Moreover, an improved block-based time-varying autoregressive model is proposed for the speech signal, which aims to reflect the underlying signal statistics more accurately on both a local and global level. Given these parametric models, their coefficients are estimated using Bayesian inference, so that the channel estimate can then be used for dereverberation. This paper presents an in-depth discussion about the applicability of these models to real speech and a real acoustic environment. Results are presented to demonstrate the performance of the Bayesian inference algorithms.

I. Introduction

Audio signals acquired in confined acoustic spaces exhibit reverberation due to multiple reflections of the sound wave from surrounding obstacles. In addition to the direct path

signal, the sensor receives time-shifted versions of the clean audio signal, leading to spectral colouration and reduced intelligibility. The reverberant signal can be modelled as a linear convolution of the source signal and the acoustic impulse response (AIR) of the room between the source and sensor. Therefore, the source signal can be enhanced by deconvolving the observations with the inverse of the degrading channel. However, in practice, neither the source nor channel are known. Since only the observed signal is available, this blind deconvolution problem is under-determined; i.e., more unknowns than observations must be estimated from a single realisation of the measurement process at each time instance. Incorporating prior knowledge of the statistical properties of the source and channel is essential for solving this problem.

Spatial diversity of acoustic channels can be constructively exploited by multiple sensor blind dereverberation techniques [1], [2] in order to obtain a clean speech estimate. However, there are numerous applications where only a single measurement of the reverberant signal is available. Single-sensor blind dereverberation is utilised in applications where numerous microphones prove infeasible or ineffective due to the physical size of arrays. Examples include hearing aids, hands-free telephony, and automatic speaker recognition.

Signal processing in acoustic environments is often approached assuming the AIR is time-invariant. This is appropriate in scenarios where the source-sensor geometry is not rapidly varying, for example, a hands-free kit in a car cabin, in which the driver and the microphone are approximately fixed relative to one another, or in a work environment where a user is seated in front of a computer terminal. However, there are many applications where the source-sensor geometry is subject to change; the wearer of a hearing-aid typically wishes to move around a room, as might users of hands-free conference telephony equipment. A speaker moving in a room at 1m/sec covers a distance of 50 mm in 50 msec. This distance might be enough for the AIR to vary sufficiently that any assumption of a time-invariant acoustic channel is no longer valid. An implicit assumption often made is that the room acoustics are time-invariant, and that it is the variable source-sensor geometry that leads to the changing AIR. However, it is not beyond possibility that the room acoustics may vary; the changing state of doors, windows, or items being moved in the room will influence the room dynamics.

Although there is some recent work dealing with time-varying acoustic channels [3], [4], generally the problem of single-channel blind dereverberation of speech from a moving speaker has received little attention from the signal processing community, in part because the case of a stationary speaker has not yet been solved satisfactorily. Nevertheless, it is still an interesting and worthwhile problem to consider.

[Fig. 1 about here.]

This paper proposes a model based framework for single-channel blind dereverberation of speech from a moving speaker by extending the work in [5]. In this approach, parametric models are used for both the source and the channel, as shown in Fig. 1. The parameters of the entire model are estimated using the Bayesian paradigm, and the source signal estimate is obtained by inverse filtering the observed signal with the estimated channel coefficients. There are two novel extensions discussed in this paper:

1) utilising a more general and flexible block-based time-varying AR (TVAR) process to model the speech signal;
2) using a linear time-varying (LTV) all-pole filter to represent the acoustic channel.

In each case, the time-varying nature of the unknown model parameters is captured by modelling them by a linear combination of known basis functions with unknown weightings as discussed in [6], [7].

Model-based approaches fundamentally rely on the availability of realistic and tractable models that reflect the underlying speech processes and acoustic systems. The choice of these models is extremely important. The underlying time-varying nature of speech signals and the rationale for the proposed model is discussed in section §II. Likewise, the proposed channel model is discussed in section §III based on observations of simulated and measured spatially-varying AIRs. The mathematical framework and methodology for parameter estimation and dereverberation is discussed in section §V. In section §VI, results using the proposed model are presented. Conclusions are drawn in section §VII.

II. SOURCE MODEL

A. Motivation

LTV all-pole filters are a popular approach for modelling the vocal tract of a speaker due to their ability to accurately model the continuous short-term spectrum of speech [8], [9]. Some sounds that are generated through a coupling between oral and nasal tracts, for example French nasals [10], must be represented by pole-zero pairs and cannot be represented by all-pole filters. Nevertheless, pole-zero speech models generally require non-linear methods for estimating their parameters [11]. Speech can be modelled as a time-varying AR (TVAR) process [11]–[14] in which the input to the all-pole filter representing the vocal tract is a white Gaussian excitation. A Q^{th}-order TVAR process is defined by

$$s(n) = -\sum_{q\in\mathcal{Q}} b_q(n)s(n-q) + e(n), \tag{1}$$

where $n \in \mathcal{N}$ is the time index over one segment of speech for N speech samples,[1] $e(n) \sim \mathcal{N}(0,\, \sigma_e^2(n))$ is the time-varying excitation with variance $\sigma_e^2(n)$, $s(n)$ is the source signal, and $\{b_q(n)\}_{q\in\mathcal{Q}}$ are the TVAR coefficients. In this framework, the problem of modelling the speech signal itself reduces to an appropriate model for the TVAR parameters, $b_q(n)$. Determining such a model is complicated, in part an open question, and is often constrained by the availability of suitable and tractable parameter estimation techniques.

Many statistical estimation methods impose stationarity on the model of the signal primarily to constructively exploit ergodicity. Since the vocal tract is continually changing with time, such a limitation results in poor modelling. In order to partially reconcile global nonstationarity while utilising the advantages of local ergodicity in estimation methods, a compromise approach is to model speech as a block-stationary process: the signal is divided into short segments where the statistics of the signal are assumed to be *locally* stationary within blocks, but *globally* time-varying, e.g.,

$$s(n) = -\sum_{q\in\mathcal{Q}} b_{qi}s(n-q) + e(n), \tag{2}$$

[1]Unless stated otherwise, the set notation $\mathcal{U} = \{1, \ldots, U\}$, where U is an integer, is used for simplicity.

where $\{b_{qi}\}_{q\in\mathcal{Q}}$ are the block stationary AR (BSAR) coefficients in block $i \in \mathcal{L}$ that are stationary within each block but vary over different blocks i.

[Fig. 2 about here.]

To illustrate the time-varying nature of speech, consider taking a sliding window of block length N over a segment of speech; the window moves by one sample in each of L steps. In each window, the Q stationary autoregressive (AR) coefficients are computed by solving the standard Yule-Walker equations [15]. The corresponding poles are the roots of the characteristic equation. For the two segments of speech shown in the grey regions in Fig. 2, the corresponding pole variations introduced by the sliding window are shown in Figs. 3a and 3b (grey dots). The poles exhibit smooth variation over these segments of speech; this characteristic of pole movements is discussed in, for example, [14]. Smooth pole variation often leads to relatively smooth parameter variation (Fig. 3c).

[Fig. 3 about here.]

Thus, the block-stationary AR model of eqn. (2) which assumes local stationarity within such segments, will not generally capture the underlying statistics of the source signal. On the other hand, the most general variation of the parameters, $b_q(n)$, in eqn. (1) is when the parameters are completely uncorrelated at each sample. In this case, each sample of the signal is represented by more than one unknown coefficient. This over-determined parameterisation results in numerical problems as there is not enough data from a single realisation of a process to allow accurate parameter estimation.

B. Basis function representation

To introduce correlation to the model, the parameters could, for instance, be represented by a random walk [16]. Alternatively, correlation is introduced by a transformation of the nonstationary signal to a space where it can be analysed as a linear time-invariant (LTI) process [6], [7], [13], [14], [17]–[19]. This corresponds to modelling the parameters as a linear combination of basis functions. To ensure that the correct model order is chosen, model order selection procedures should be implemented: [19] proposes such an algorithm

based on the discrete Karhunen-Loève transform.

Ideally, the pole locations rather than the parameter variation are represented as a function of time by a parametric model. However, this is difficult as the relationship between poles and parameters is non-linear and a closed-form expression for the pole positions for high-order models cannot be derived. If the TVAR coefficients can be represented by a linear combination of basis functions, eqn. (1) can be formulated as [7], [13]:

$$s(n) = -\sum_{q\in\mathcal{Q}} \underbrace{\left\{\sum_{k\in\mathcal{F}} b_{qk} f_k(n-q)\right\}}_{b_q(n)} s(n-q) + e(n), \tag{3}$$

where F is the number of basis functions, $\mathbf{b} = \{b_{qk} : q \in \mathcal{Q};\, k \in \mathcal{F}\}$ are the *unknown* time-invariant basis coefficients, and $\{f_k(n)\}_{k\in\mathcal{F}}$ are the *known* time-varying basis functions.

To demonstrate that the speech pole movements can be approximated by the model in eqn. (3), a least-squares (LS) fit to the AR parameters corresponding to the speech pole movements in Figs. 3a and 3b is performed using the trigonometric Fourier basis set $\{\sin(n\omega_0 t),\, \cos(n\omega_0 t)\}_{n=0}^{2}$ with fundamental frequency $\omega_0 = 2\pi\frac{5}{9}$ rad/sec. Due to the linearity of the source model in eqn. (3), the basis coefficients, $\mathbf{b}$, are obtained as the linear LS estimate [15]. The full TVAR coefficients, $\{b_q(n)\}$, are then estimated by multiplication of the basis functions with the linear LS estimate of the basis coefficients. The estimates of the TVAR parameters are depicted in Figs. 3a and 3b in black dots, and show a good match to the actual poles (Fig. 3d). This and the results in [7], [13], [14], [17], [18] lead to the conclusion that a model based on the transformation from a LTV process to a LTI process through a set of basis functions can capture appropriately the time-variation of short segments of speech.

C. Choice of basis functions

As the basis functions span the vector space to which the source signal is mapped, their choice is essential. Unfortunately, no general rules for choosing these functions exist. The choice of basis is therefore dependent on the prior belief of the variation of the parameters. Amongst the wide range of basis functions that have been investigated [13], [14], [18], [20], standard choices include Fourier functions, Legendre polynomials, and discrete prolate spheroidal sequences (DPSS). These classes tend to assume smooth parameter behaviour

and respond to abrupt changes as a low-pass filter [14]. Hence, for abrupt changes in the source signal, the parameters are not modelled correctly. Discontinuous basis like the step function that is used for BSAR processes capture abrupt changes well, but cannot handle smooth variations [14]. Modelling rapid parameter variation is theoretically possible by utilising an infinite number of basis functions. However, this would again reduce the model to a time-varying process with uncorrelated parameters as described above, leading to over-parameterised coefficients since the model would have as many degrees of freedom as the signal itself [14], [19].

A comparison of the performance of different basis sets for speech is beyond the scope of this paper, although a comparison of signal modelling using Fourier, Legendre and DPSS basis sets is detailed in Charbonnier *et al.* [18]. In this paper, it is assumed for simplicity that the true speech parameters can be approximated by sinusoidal functions (Fourier basis), since these are seen to be a good model the source parameter variations (grey line) as depicted in Fig. 3c.

The difficulty of abrupt parameter variations is seen in Fig. 3a, where some of the speech poles evolve towards the origin and then abruptly jump away from it. Since the frequency response of poles approaching the origin becomes increasingly flat, this pole behaviour corresponds to a birth-death process. This effect does not occur for the same experiment using a lower order due to a more parsimonious representation. In other words, the death and birth of poles is an artifact introduced through the over-parameterisation of the model. Ideally, the system should have a time-varying model order so as to capture poles that contribute to the frequency response of the speech signal, and adjust the model order when poles become redundant. Thus, the model order, Q, and the block-length, N, (see eqn. (4) in the next section) are in principle also random variables and could be allowed to vary with the block index. While this would capture any birth or deaths of poles, the estimation techniques required such as reversible-jump Markov chain Monte Carlo (MCMC) methods greatly increase the computational burden and implementation complexity.

D. Block-based time-varying approach

[Fig. 4 about here.]

An alternative approach to addressing the issue of abrupt parameter variations while using a limited set of basis functions is proposed, and relies on a block-based time-varying model. Here, the signal is segmented into shorter blocks that are modelled locally as well as globally time-varying. Instead of utilising one set of parameters coping with rapid global variation, several sets of parameters are introduced that capture the local variation within each block. For sufficiently short blocks, the time variation of the signal will be smooth and parameters can be estimated accurately using a standard choice of basis functions.

This model thus attempts to incorporate the time-varying nature of the signal both locally as well as globally. The advantages and disadvantages of stationarity and nonstationarity on a local and global level are outlined in TABLE I.

In the block-based TVAR model, the source signal is expressed for a block of data, indexed by i and of length $N_i = T_{i+1} - T_i$, for samples $n \in \mathcal{T}_i = \{T_i, \ldots, T_{i+1} - 1\}$ as:

$$s(n) = -\sum_{q \in \mathcal{Q}} \underbrace{\left\{ \sum_{k \in \mathcal{F}} b_{iqk}\, f_k(n - T_i + Q - q) \right\}}_{b_q(n),\, n \in \mathcal{T}_i} s(n-q) + e(n), \tag{4}$$

where $e(n) \sim \mathcal{N}\left(0,\, \sigma^2_{e,i}\right)$ has variance $\sigma^2_{e,i}$ and the block boundaries are specified by T_i and T_{i+1} in block $i \in \mathcal{L}$. This model is illustrated in Fig. 4, and reduces to the TVAR model (eqn. (1)) in the case of a single block. Unlike the examples presented in section §II-A and Fig. 3, the blocks in this model are *non-overlapping*. Note this model implicitly assumes unvoiced speech segments as it uses a white excitation. An issue for further research is whether the model also works effectively for voiced speech.

[TABLE 1 about here.]

III. Channel Model

There are many different techniques for modelling an acoustic impulse response (AIR) and, in general, each model applies to a different frequency range of the audible spectrum. The acoustic response of a room, $h(t)$, takes the general form:

$$h(t) = \begin{cases} 0 & \text{for } t < 0 \\ \sum_n \tilde{A}_n e^{-\tilde{\delta}_n t} \cos\left(\tilde{\omega}_n t + \tilde{\theta}_n\right) & \text{for } t \geq 0 \end{cases} \tag{5}$$

where the coefficients $\tilde{A}_n$ implicitly contain the location of the source and observer, $\tilde{\delta}_n$, $\tilde{\omega}_n$, and $\tilde{\theta}_n$ are the damping constant, undamped natural frequency, and phase terms respectively. Although this general parametric model completely characterises the acoustic impulse response, it is intractable for many estimation problems and does not lead to an analytical solution in this blind dereverberation framework. The problem from a signal processing perspective is that there is no practical model for the entire audible frequency range [21].

A. Characteristics of room acoustics

Generally, the audible spectrum can be divided into four regions. Consider a single-tone source with frequency f generated in a room with dimensions $2.78 \times 4.68 \times 3.2\,\mathrm{m}$ and reverberation time of $T_{60} = 0.23$ sec. (as used section §III-D).

Very Low Frequencies: For $f < f_w = \frac{c}{2L}$, where c is the speed of sound, and L is the largest dimension of the acoustic environment, there is no resonant support. Typically, f_w is around 35 Hz for this room.

Wave Acoustics: This region corresponds to frequencies for which the source's wavelength is comparable to the room dimensions. It spans the lowest resonant mode, given by $\approx f_w$, to the Schroeder frequency $f_g \approx 2000\sqrt{\frac{T_{60}}{V}}$ (Hz) where V is the volume of the room. Distinct resonants occur in which the quality-factor (Q-factor) is sufficiently large that the average spacing of resonant frequencies is substantially larger than the average *half-width* of the resonant mode. For this room, distinct resonances occur between 35 Hz and 149 Hz.

Very low frequency regions and wave acoustics are generally irrelevant for speech dereverberation as electro-acoustic systems have a limited bandwidth at low frequencies. Analytical tools are thus utilised only for the following regions:

High Sound Frequencies: Above f_g, there is such a strong model overlap that the concept of a resonant mode becomes meaningless. However, below a frequency of around $4f_g$, the wavelengths are too long for the application of *geometric acoustics* discussed below. Thus, a statistical treatment is generally employed. For the room above, statistical theory would be relevant between 149 Hz and 595 Hz.

Geometrical Acoustics: Above $4f_g$, *geometrical room acoustics* apply and assumes the limiting case of vanishingly small wavelengths. This assumption is valid if the dimensions

of the room and its walls are large compared with the wavelength of sound: this condition is met for a wide-range of audio frequencies in standard rooms. In this frequency range, specular reflections and the *sound ray* approach to acoustics prevail. Geometrical acoustics usually neglect wave related effects such as diffraction and interference. The image method [22] for simulated AIRs is valid only in this frequency range.

B. Pole-zero and all-zero models

The solution of the acoustic wave equation indicates that a room transfer function can be expressed by a rational expression, and therefore can be modelled by a conventional pole-zero model. Mourjopoulos and Paraskevas [23] discuss pole-zero modelling of room transfer functions (RTFs), and the model has often been used in the literature. From a physical point of view, poles represent resonances, and zeros represent time delays and anti-resonances. Another commonly used model is the all-zero model. There are several main limitations of finite impulse response (FIR) filters imposed by the nature of room acoustics [23], [24]. First, AIRs are, in general, very long and an all-zero filter typically requires $n_s = T_{60} f_s$ coefficients where f_s is the sampling frequency. For example, if $T_{60} = 0.5$ seconds and $f_s = 10$ kHz, the all-zero filter requires $n_s = 5000$ coefficients. Secondly, the resulting FIR filter may be effective only for a limited spatial combination of source and receiver positions, as all-zero models lead to large variations in the room transfer function for small changes in source–observer positions [23], [24]. A further disadvantage of the pole-zero and all-zero models is that estimation of the zeros requires solving a set of non-linear equations.

C. All-pole models and basis function representation

As an alternative, the all-pole model for approximating rational transfer functions is widely used in many fields. It is claimed that typical all-pole model orders required for approximating room transfer functions are in the range $50 \leq P \leq 500$ [23], although this depends on the frequency range of the acoustic spectrum considered. A significant advantage of the all-pole model over the all-zero model is its lower sensitivity to changes in source and observer positions. Mourjopoulos and Paraskevas [23] conclude that in many signal processing applications dealing with room acoustics, it may be both sufficient and more efficient to manipulate all-pole

model coefficients rather than high-order all-zero models. All-pole models are particularly useful for modelling resonances in the wave acoustics and high sound frequency regions.

If the source signal, $s(n)$, is filtered through an AIR modelled by an all-pole filter of order P, the observed signal, $x(n)$, received at the microphone, can be expressed as

$$x(n) = -\sum_{p\in\mathcal{P}} a_p(n)x(n-p) + s(n), \tag{6}$$

where $\{a_p(n)\}_{p\in\mathcal{P}}$ are the time-varying all-pole channel coefficients. Following the reasoning in section §II-A, similar to eqns. (1) and (3), the channel coefficients are represented by a linear combination of basis functions, and hence the time-varying channel is formulated as

$$x(n) = -\sum_{p\in\mathcal{P}} \underbrace{\left\{\sum_{\ell\in\mathcal{G}} a_{p\ell} g_\ell(n-p)\right\}}_{a_p(n)} x(n-p) + s(n), \tag{7}$$

where $\{a_{p\ell} : p \in \mathcal{P};\ \ell \in \mathcal{G}\}$ are the G unknown time-invariant basis coefficients, $\{g_\ell(n)\}_{\ell\in\mathcal{G}}$ are the known time-varying basis functions. Note that eqn. (7) applies over all blocks, i.e., the channel model is *not* block-based.

D. Choice of basis functions

[Fig. 5 about here.]

[Fig. 6 about here.]

In order to select an appropriate set of basis functions for modelling the variation of the all-pole coefficients, the spatially-varying nature of AIRs is briefly investigated. Simulated and measured AIRs are obtained for the acoustic set-up illustrated in Fig. 5 for a small office of size $2.78 \times 4.68 \times 3.2$m (length $\times$ width $\times$ height). An acoustic source remains fixed while the microphone sensor is moved from its initial position in 2 mm increments. This experimental set-up mimics the spatially-varying nature of the AIR for non-stationary sources.

The simulated AIRs are generated using the image method [22] with the reflection coefficient chosen to give a reverberation time of $T_{60} = 0.23$ seconds. This choice corresponds to the measured reverberation time of the real office. As the image model assumes geometric

room acoustics, the simulated responses only apply above four times the Schroeder frequency, f_g, as discussed in section §III-A, and in this case $4f_g = 595$ Hz. Using the simulated AIRs, the RTF is modelled in the frequency range between 600 Hz to 1200 Hz by a 16^{th}-order sub-band AR model [25]. The variation of the resulting pole positions from the initial sensor position to a final offset of 400 mm is plotted in Fig. 6a. The results indicate smooth pole variation and, consequently, the TVAR parameters of the AIR vary relatively smoothly with sensor spatial displacement. This can be confirmed by measures of the changes in the AIR, e.g., normalised projection misalignment.

For verification of these results using real data, 910 AIRs were measured in a real office by moving a 26-microphone linear array in small increments over a distance of 70 mm. To obtain comparable results to the simulated data, the pole variations are again acquired by modelling the RTF as a 16^{th}-order AR sub-band model in the range 600 Hz to 1200 Hz. The poles for real AIRs are subject to larger variation than those for the simulated AIRs, they cover a wider region within the unit circle, and intersect the trajectories of neighbouring poles. To avoid cluttered pole trajectory plots, only a subset of the pole variations from the microphone array for several microphones (labelled mics. 7 and 8) are displayed in Figs. 6c and 6d. This corresponds to offsets from 432mm to 502mm for mic. 7 and from 504mm to 574mm for mic. 8. For comparison with equivalent results for simulated data see Fig. 6b. The pole variations from the measured data clearly exhibit reasonably smooth trajectories, validating the simulated results.

An in-depth discussion of the variability of room acoustics is beyond the scope of this paper, and requires considerably more investigation than the results presented in this section. Nonetheless, based on the results presented in Fig. 6, it is concluded that basis functions could be used for capturing the smooth variations of the poles and parameters in the model of the AIR. Following the discussion in section §II-C, Fourier basis functions will therefore be utilised in the following.

E. Modelling issues

Single-channel blind dereverberation is a notoriously difficult and challenging problem. In the approach used here,[2] the acoustic channel is blindly estimated from the reverberant signal, and then used in deconvolution to obtain the anechoic signal. There are a number of problems encountered when dealing with acoustic impulse responses (AIRs) [26].

High number of channel parameters: The length of AIRs, as discussed in section §III-B, make estimates difficult.

Nonminimum-phase responses: AIRs are often nonminimum-phase, and leads to difficulties with channel modelling and inversion. The nonminimum-phase contribution to the perception of reverberation is significant [27], [28].

Robustness to estimation error: Any small error in an AIR estimate leads to a significant error in the inverse of the AIR. Thus, inversion can increase distortion in the enhanced signal compared to the reverberant signal. Any deviation from the true AIR means that attempts to equalise high-Q resonances can still leave high-Q resonances in the equalised response degrading the intelligibility of the restored signal.

Variation of inverse of AIR: Similarly, while a small change in source-sensor geometry might give rise to a small change in the AIR as shown previously, the corresponding changes in the inverse of an AIR can sometimes be large.

Since the proposed channel estimation techniques and source recovery method implicitly uses inverse-filtering methods, these issues are particularly pertinent. Some of these problems can be alleviated by neither attempting to process the full frequency range of the source, nor attempting to invert the *full-band* RTF using a single filter. In problems with long channels, it is better to utilise sub-band methods that attempt to enhance the reverberant signal by inverting the channel response over a number of separate frequency ranges. Modelling each frequency band independently can lead to a parsimonious approximation of the RTF, lower model orders, and an overall reduction in the total number of parameters needed to approximate the acoustic channel [25]. Moreover, there may be only a few bands that have high-Q resonances

[2]Another distinct approach to blind dereverberation is as an optimal filtering formulation in which estimates of the unknown source signal are estimated directly from the reverberant data [3].

which need careful equalisation, whereas other frequency bands have lower Q factors, so less care is required.

An additional advantage of using sub-band models is that sub-bands possessing minimum-phase characteristics can be inverted, despite the AIRs being nonminimum-phase over the full frequency range. Hence, in the case of a nonminimum-phase response, where a causal inverse does not exist, methods for detecting and equalising the minimum-phase sub-bands should be developed: this follows the approaches in [29], [30]. Details of the sub-band methodology are discussed in [25] and can be incorporated into the framework proposed in this paper.

IV. Source and Channel Identifiability

Single-channel blind dereverberation is an inherently under-determined problem. For example, if both source and channel are modelled as stationary AR processes, the observed signal is also a stationary AR process. Consequently, it is not possible to attribute a particular pole estimated from the observed signal to either the source or channel: there is an identifiability ambiguity. Source-channel ambiguities can be avoided by, for example, modelling the acoustic source as a TVAR process, and the channel by a FIR filter. The observed signal is then a time-varying ARMA process, in which the poles belong to the source model and zeros to the channel. Thus, there appears to be no ambiguity in distinguishing between the parameters associated with each. This model is used in [3] for the case of separating and recovering convolutively mixed signals. However, this is not always a realistic model, as it cannot be ascertained that the source only has poles and no zeros, and the channel only has zeros, and no poles.

In an earlier approach to single-channel blind dereverberation focusing on stationary speakers [5], the locally-stationary nature of the source and the *assumed* time-invariance of the channel were utilised to provide sufficient information to distinguish between the two models. In this approach it was argued that the statistics of speech signals remain quasi-stationary for around 20-50 msec. The source signal is modelled by a BSAR process, while the AIR is modelled by a LTI all-pole filter. These models allow the acoustic channel to be uniquely identified up to a scaling ambiguity, since essentially any common poles estimated from different blocks of the observed data must belong to the channel.

As discussed in section §II, this paper presents an improved system model by using a block-based TVAR model and a time-varying all-pole channel filter. The question now is whether there are identifiability ambiguities in such a block-based TVAR-TVAR model. Although this question is not comprehensively addressed here, the following are contributing factors to the identifiability issue.

1) While the cascade of two LTI systems commute, LTV systems, in general, do not. Since the source and channel have different time-varying characteristics, the system is likely to have a unique source-channel decomposition.
2) Consequently, a block-based source model, whether BSAR or block-based TVAR, tends to reduce ambiguities in identifying the source and channel – since the channel parameters are stationary over all data samples, whereas the source parameters vary on a block basis.
3) If the chosen basis functions do not allow accurate tracking of the TVAR parameters, the models will not fit the data accurately, and the estimates will be poor.
4) The source signal needs to be spectrally rich in order to provide sufficient energy to 'illuminate' the channel, such that there is enough information in the observations for identifiability.

To illustrate this last point, consider taking a *temporal average* of the source signal. If the source contains relatively little energy at spectral frequencies in which there is significant channel information, such as key resonances, the channel estimates in that spectral region will be poor. Consequently, this suggests that the poles in the source and channel models need to lie in a region where they contribute sufficiently to the spectral content of the observed signal. Thus, referring back to section §II-D, this indicates that poles which undergo a birth and death procedure will be difficult to identify. Further discussion and results regarding this are given in section §VI.

V. Bayesian Blind Model Parameter Estimation

The observed reverberant signal, $x(n)$, is given by eqn. (7). If the channel parameters $\{a_{p\ell}\}$ can be estimated, the source signal, $s(n)$, can easily be recovered through a rearrangement of eqn. (7), in what is essentially an inverse filtering operation. However, finding the channel

parameters requires finding the source parameters $\{b_{iqk}\}$ in eqn. (4) as well. Since the source excitation is white Gaussian, the estimation of all the system model parameters can be achieved using maximum-likelihood methods such as the Expectation-Maximisation algorithm [31].

In this paper, Bayesian inference and associated numerical optimisation methods are used for this parameter estimation. Bayes's rule provides a learning procedure where knowledge of the system is inferred from prior belief and updated through new data. Consider a data model, $\mathcal{M}$, with unknown parameters, $\boldsymbol{\theta}_{\mathcal{M}}$, for the N samples of observed data, $\mathbf{x} = \{x(n),\, n \in \mathcal{N}\}$. The posterior probability, $p\left(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{M}\right)$, for the unknown parameters is defined by Bayes's theorem as

$$p\left(\boldsymbol{\theta}_{\mathcal{M}} \mid \mathbf{x}, \mathcal{M}\right) = \frac{p\left(\mathbf{x} \mid \boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}\right) p\left(\boldsymbol{\theta}_{\mathcal{M}} \mid \mathcal{M}\right)}{p\left(\mathbf{x} \mid \mathcal{M}\right)}, \tag{8}$$

where $p\left(\mathbf{x} \mid \boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}\right)$ is the likelihood, $p\left(\boldsymbol{\theta}_{\mathcal{M}} \mid \mathcal{M}\right)$ is the prior distribution on $\boldsymbol{\theta}_{\mathcal{M}}$. The term $p\left(\mathbf{x} \mid \mathcal{M}\right)$ is called the evidence, and is usually regarded as a normalising constant.

Given the likelihood function and the prior distributions, Bayesian methods aim to estimate the unknown parameters from the posterior distribution. Although deterministic optimisation methods for determining the maximum marginal *a posteriori* (MMAP) estimate could be used to directly locate the mode of the posterior, this becomes unreliable for high-dimensional multi-modal distributions. Thus, iterative stochastic sampling schemes are used: MCMC methods are based on constructing a Markov chain that has the desired distribution as its invariant distribution. In the following, the observation likelihood and prior distributions are defined, and the Gibbs sampler introduced. The posterior density for the channel parameters given the observations, as well as the conditional distributions required for Gibbs sampling are outlined.

A. Likelihood for the source signal and observations

1) Source Model: Rewriting a vector of excitation samples, $e(n)$, in eqn. (4) in block, i, for $n \in \mathcal{T}_i = \{T_i, \ldots, T_{i+1} - 1\}$,

$$\mathbf{e}_i = \mathbf{B}_{i,\text{blk}}\, \mathbf{s}_i + \mathbf{B}_{i,\text{ini}}\, \mathbf{s}_{i-1,Q} \tag{9a}$$

$$= \underbrace{\begin{bmatrix}\mathbf{B}_{i,\text{ini}} & \mathbf{B}_{i,\text{blk}}\end{bmatrix}}_{\mathbf{B}_i \in \mathbb{R}^{N_i \times (N_i+Q)}} \underbrace{\begin{bmatrix}\mathbf{s}_{i-1,Q} \\ \mathbf{s}_i\end{bmatrix}}_{\hat{\mathbf{s}}_i \in \mathbb{R}^{(N_i+Q)\times 1}} = \mathbf{B}_i\, \hat{\mathbf{s}}_i, \tag{9b}$$

where the error residual in block i, $\mathbf{e}_i = \begin{bmatrix} e(T_i) & \cdots & e(T_{i+1}-1)\end{bmatrix}^T$, is a $N_i \times 1$ vector with $N_i = T_{i+1} - T_i$ samples per block, T_i and T_{i+1} denotes the first samples of the current and next block, respectively. The $N_i \times 1$ vector containing the source signal samples in block i is $\mathbf{s}_i = \begin{bmatrix} s(T_i) & \cdots & s(T_{i+1}-1)\end{bmatrix}^T$, and the $Q \times 1$ vector containing the last Q samples of the data in the previous block, $i-1$, is $\mathbf{s}_{i-1,Q} = \begin{bmatrix} s(T_i - Q) & \cdots & s(T_i - 1)\end{bmatrix}^T$. This vector, $\mathbf{s}_{i-1,Q}$, is referred to as the initial conditions for block i. The $N_i \times N_i$ matrix, $\mathbf{B}_{i,\text{blk}}$, of the TVAR coefficients in block i is appropriately defined and takes the form

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ b_1(T_i+1) & 1 & 0 & \dots & 0 & 0 \\ b_2(T_i+2) & b_1(T_i+2) & 1 & 0 & \dots & 0 \\ \ddots & & & \ddots & & \\ \cdots & 0 & b_Q(T_{i+1}-1) & \dots & b_1(T_{i+1}-1) & 1 \end{bmatrix}.$$

The $N_i \times Q$ matrix containing the initial conditions of the TVAR coefficients is

$$\mathbf{B}_{i,\text{ini}} = \begin{bmatrix} b_Q(T_i) & b_{Q-1}(T_i) & \cdots & b_1(T_i) \\ 0 & b_Q(T_i+1) & \cdots & b_2(T_i+1) \\ 0 & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & b_Q(T_i+Q-1) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Assuming $e(n)$ in block i is stationary white Gaussian noise (WGN), then applying the probability transformation $\mathbf{e}_i \mapsto \mathbf{s}_i$, the likelihood of the source signal in block i is found as

$$p_{\mathbf{s}_i}\left(\mathbf{s}_i \mid \mathbf{b}_i, \sigma^2_{e,i}, \mathbf{s}_{i-1}\right) = \frac{1}{\left(2\pi\sigma^2_{e,i}\right)^{\frac{N_i}{2}}} \exp\left\{-\frac{1}{2\sigma^2_{e,i}} \left\|\mathbf{B}_i\, \hat{\mathbf{s}}_i\right\|^2\right\}. \tag{10}$$

Applying the probability chain rule to all the data across multiple blocks denoted by $\mathbf{s} =$

$\begin{bmatrix}\mathbf{s}_0^T & \dots & \mathbf{s}_L^T\end{bmatrix}^T$, the likelihood for all source data across L blocks of data is given by

$$\begin{aligned} p_{\mathbf{S}}\left(\mathbf{s} \mid \mathbf{b}, \boldsymbol{\sigma}_e\right) &= p_{\mathbf{S}_0}\left(\mathbf{s}_0\right) \prod_{i \in \mathcal{L}} p_{\mathbf{S}_i}\left(\mathbf{s}_i \mid \mathbf{b}_i, \sigma_{e,i}^2, \mathbf{s}_{i-1}\right) \\ &= p_{\mathbf{S}_0}\left(\mathbf{s}_0\right) \prod_{i \in \mathcal{L}} \frac{1}{\left(2\pi\sigma_{e,i}^2\right)^{\frac{N_i}{2}}} \exp\left\{-\frac{1}{2\sigma_{e,i}^2}\left\|\mathbf{B}_i\,\hat{\mathbf{s}}_i\right\|^2\right\}, \end{aligned} \tag{11}$$

where $\mathbf{s}_0$ is the data upon which the first block is conditional. The term $p_{\mathbf{S}_i}\left(\mathbf{s}_i \mid \mathbf{b}_i, \sigma_{e,i}^2, \mathbf{s}_{i-1}\right)$ represent the probability density function (pdf) for the signal in the i^{th} block and is conditional on values outside that block. The unconditional pdf $p_{\mathbf{S}_0}\left(\mathbf{s}_0\right)$ representing the 'initial' data for the first block takes on a more complex form as discussed in [32]. For a large amount of data, it is reasonable and often assumed to be constant, so that this term can be omitted from eqn. (11).

A linear-in-the-parameters (LITP) representation is obtained for the model by writing eqn. (4) in matrix-vector form as

$$\mathbf{s}_i = -\sum_{q \in \mathcal{Q}} \mathbf{S}_{i,q}\mathbf{F}_{i,q}\mathbf{b}_{i,q} + \mathbf{e}_i, \tag{12}$$

where the $N_i \times 1$ vector of source samples is $\mathbf{s}_{i,q} = \begin{bmatrix}s(T_i - q) & \cdots & s(T_{i+1} - 1 - q)\end{bmatrix}^T$ and the $N_i \times N_i$ matrix $\mathbf{S}_{i,q} = \text{diag}\left[\mathbf{s}_{i,q}\right]$, where $\text{diag}\left[\cdot\right]$ denotes a diagonal matrix. Furthermore, $\mathbf{F}_{i,q}$ is a $N_i \times F$ matrix whose columns contain the F basis functions, such that the (j,k)-th element of $\mathbf{F}_{i,q}$ is $[\mathbf{F}_{i,q}]_{jk} = f_k(j+Q-q)$. Defining the $N_i \times FQ$ matrix $\mathbf{U}_i \triangleq \begin{bmatrix}\mathbf{U}_{i,1} & \dots & \mathbf{U}_{i,Q}\end{bmatrix}$, where $\mathbf{U}_{i,q} = \mathbf{S}_q\mathbf{F}_q$, and the $FQ \times 1$ vector $\mathbf{b}_i \triangleq \begin{bmatrix}\mathbf{b}_{i,1}^T & \dots & \mathbf{b}_{i,Q}^T\end{bmatrix}^T$, where $\left[\mathbf{b}_{i,q}\right]_k = b_{iqk}$, eqn. (12) can be written as

$$\mathbf{e}_i = \mathbf{s}_i + \mathbf{U}_i\mathbf{b}_i. \tag{13}$$

Therefore, the source likelihood in eqn. (11) is equivalent to

$$p_{\mathbf{S}}\left(\mathbf{s} \mid \mathbf{b}, \boldsymbol{\sigma}_e\right) \approx \prod_{i \in \mathcal{L}} \frac{1}{\left(2\pi\sigma_{e,i}^2\right)^{\frac{N_i}{2}}} \exp\left\{-\frac{1}{2\sigma_{e,i}^2}\left\|\mathbf{s}_i + \mathbf{U}_i\mathbf{b}_i\right\|^2\right\}. \tag{14}$$

2) Channel Model: In a similar construction to eqn. (9b), eqn. (7) can be written as,

$$\mathbf{s} = \mathbf{A}_{\text{blk}}\,\hat{\mathbf{x}} + \mathbf{A}_{\text{ini}}\,\mathbf{x}_{\text{ini}} = \underbrace{\begin{bmatrix}\mathbf{A}_{\text{ini}} & \mathbf{A}_{\text{blk}}\end{bmatrix}}_{\mathbf{A}\in\mathbb{R}^{N_x\times(N_x+P)}} \underbrace{\begin{bmatrix}\mathbf{x}_{\text{ini}} \\ \hat{\mathbf{x}}\end{bmatrix}}_{\mathbf{x}\in\mathbb{R}^{(N_x+P)\times 1}} = \mathbf{A}\mathbf{x}, \tag{15}$$

where $n = \{P, \ldots, N-1\}$, N is the total number of output samples,[3] and the actual number of observations is $N_x = N - P$. Thus, let the $N_x \times 1$ vector of the observations be $\hat{\mathbf{x}} = \begin{bmatrix} x(P) & \cdots & x(N-1)\end{bmatrix}^T$, and assume the $P \times 1$ vector of initial conditions in $\mathbf{x}_{\text{ini}} = \begin{bmatrix} x(0) & \cdots & x(P-1)\end{bmatrix}^T$ is known. The $N_x \times 1$ vector of source samples is $\mathbf{s} = \begin{bmatrix} s(P) & \cdots & s(N-1)\end{bmatrix}^T$. The $N_x \times N_x$ matrix containing the TVAR channel coefficients is

$$\mathbf{A}_{\text{blk}} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ a_1(P+1) & 1 & 0 & \ldots & 0 \\ & \ddots & & \ddots & \\ \cdots & a_P(N-1) & \cdots & a_1(N-1) & 1 \end{bmatrix},$$

and the $N_x \times P$ matrix containing the initial conditions of the TVAR channel coefficients is

$$\mathbf{A}_{\text{ini}} = \begin{bmatrix} a_P(P) & a_{P-1}(P) & \cdots & a_1(P) \\ 0 & a_P(P+1) & \cdots & a_2(P+1) \\ 0 & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & a_P(2P-1) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Applying the probability transformation $\mathbf{s} \mapsto \mathbf{x}$ to eqn. (15) and using eqns. (11) and (14), the likelihood of the observations given the system parameters becomes

$$\begin{aligned} &p_{\mathbf{X}}\left(\mathbf{x} \,|\, \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e^2\right) \\ &= p_{\mathbf{x}_{\text{ini}}}\left(\mathbf{x}_{\text{ini}}\right) \left[\prod_{i\in\mathcal{L}} \frac{1}{\left(2\pi\sigma_{e,i}^2\right)^{\frac{N_i}{2}}} \exp\left\{-\frac{1}{2\sigma_{e,i}^2}\left\|\mathbf{B}_i\,\hat{\mathbf{s}}_i\right\|^2\right\}\right]_{\mathbf{s}=\mathbf{A}\mathbf{x}} \end{aligned} \tag{16}$$

[3]Recall that the source is block-based, whilst the channel model is defined over all the data.

$$= p_{\mathbf{X}_{\text{ini}}}(\mathbf{x}_{\text{ini}}) \left[\prod_{i \in \mathcal{L}} \frac{1}{\left(2\pi\sigma_{e,i}^2\right)^{\frac{N_i}{2}}} \exp\left\{ -\frac{1}{2\sigma_{e,i}^2} \left\| \mathbf{s}_i + \mathbf{U}_i \mathbf{b}_i \right\|^2 \right\} \right]_{\mathbf{s}=\mathbf{A}\mathbf{x}}, \tag{17}$$

where the vectors $\{\mathbf{s}_i\}$, $\{\hat{\mathbf{s}}_i\}$ and matrices $\{\mathbf{U}_i\}$ are functions of the channel parameters and observations, as dictated through the relationship $\mathbf{s} = \mathbf{A}\mathbf{x}$. Again, assuming $p_{\mathbf{X}_{\text{ini}}}(\mathbf{x}_{\text{ini}}) \approxeq$ const., the initial terms can be omitted from the observation likelihood in eqn. (17).

B. Prior distributions of source, channel, and error residual

A prior reflects the knowledge of the parameters before the data is observed. By means of prior densities, the posterior can be manipulated by inferring any required statistic, leading to a fully interpretable probability density function. If no prior knowledge is available, the prior pdf should be broad and flat compared to the likelihood. Such priors are known as non-informative and "convey ignorance of the values of the parameters before observing the data" [31].

Since the terms in the likelihood for AR parameters are usually in the form of a Gaussian distribution [32], and in order to obtain analytically tractable results, Gaussian priors are imposed on the channel and source parameters, i.e., $p(\mathbf{a} \mid \sigma_{\mathbf{a}}^2) = \mathcal{N}\left(\mathbf{a} \,\middle|\, \mathbf{0},\, \sigma_{\mathbf{a}}^2 \mathbf{I}_P\right)$ and $p\left(\mathbf{b}_i \mid \sigma_{\mathbf{b}_i}^2\right) = \mathcal{N}\left(\mathbf{b}_i \,\middle|\, \mathbf{0},\, \sigma_{\mathbf{b}_i}^2 \mathbf{I}_Q\right)$, where $\mathcal{N}\left(x \,\middle|\, \cdot,\, \cdot\right)$ denotes a Gaussian pdf and $\mathbf{I}_K$ is the identity matrix of size $K \times K$.

A standard prior for scale parameters, such as variances, is the inverse-Gamma density.[4] The prior distributions on the error residual variance as well as the hyperparameters of the channel and source coefficients are therefore assigned as $p\left(\sigma_{e,i}^2 \,\middle|\, \alpha_{e,i}, \beta_{e,i}\right) = \mathcal{IG}\left(\sigma_{e,i}^2 \,\middle|\, \alpha_{e,i},\, \beta_{e,i}\right)$ for the error residual variance, $p\left(\sigma_{\mathbf{b}_i}^2 \,\middle|\, \alpha_{\mathbf{b}_i}, \beta_{\mathbf{b}_i}\right) = \mathcal{IG}\left(\sigma_{\mathbf{b}_i}^2 \,\middle|\, \alpha_{\mathbf{b}_i},\, \beta_{\mathbf{b}_i}\right)$ and $p(\sigma_{\mathbf{a}}^2 \mid \alpha_{\mathbf{a}}, \beta_{\mathbf{a}}) = \mathcal{IG}\left(\sigma_{\mathbf{a}}^2 \,\middle|\, \alpha_{\mathbf{a}},\, \beta_{\mathbf{a}}\right)$ for the hyperparameters on the source and channel respectively.

C. Posterior distribution of the channel parameters

The joint-posterior pdf is found using Bayes's theorem:

$$p(\mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e \mid \mathbf{x}, \boldsymbol{\phi}) \propto p(\mathbf{x} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e) \cdot p(\mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e \mid \boldsymbol{\phi}) \tag{18}$$

[4]Inverse-Gamma pdf is: $\mathcal{IG}\left(x \,\middle|\, \alpha,\, \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left\{-\frac{\beta}{x}\right\}$.

where,

$$p\left(\mathbf{a},\, \mathbf{b},\, \boldsymbol{\sigma}_e \mid \boldsymbol{\phi}\right) = p\left(\mathbf{a} \mid \sigma^2_{\mathbf{a}}\right)\, p\left(\sigma^2_{\mathbf{a}} \,\middle|\, \alpha_{\mathbf{a}},\, \beta_{\mathbf{a}}\right) \times \prod_{i\in\mathcal{L}} p\left(\mathbf{b}_i \mid \sigma^2_{\mathbf{b}_i}\right)\, p\left(\sigma^2_{\mathbf{b}_i} \,\middle|\, \alpha_{\mathbf{b}_i},\, \beta_{\mathbf{b}_i}\right)\, p\left(\sigma^2_{e,i} \,\middle|\, \alpha_{e,i},\, \beta_{e,i}\right) \tag{19}$$

assuming the system parameters $\{\mathbf{a},\, \mathbf{b},\, \boldsymbol{\sigma}_e\}$ are independent. The set $\boldsymbol{\phi} \triangleq \left\{\sigma^2_{\mathbf{a}},\, \alpha_{\mathbf{a}},\, \beta_{\mathbf{a}},\, \left\{\sigma^2_{\mathbf{b}_i},\, \alpha_{\mathbf{b}_i},\, \beta_{\mathbf{b}_i},\, \alpha_{e,i},\, \beta_{e,i}\right\}_{i\in\mathcal{L}}\right\}$ contains the hyperparameters, $\sigma^2_{\{\mathbf{a},\mathbf{b}_i\}}$ and hyper-hyperparameters, $\{\alpha_{\{\mathbf{a},\mathbf{b}_i,e_i\}},\, \beta_{\{\mathbf{a},\mathbf{b}_i,e_i\}}\}$ on the channel and source coefficients, and the error residual variance.

Ideally, from eqns. (17) and (18), the nuisance parameters $\mathbf{b}$ and $\boldsymbol{\sigma}_e$ should be marginalised out to form the marginal *a posteriori* pdf. This is derived, as shown in Appendix A, as:

$$p\left(\mathbf{a} \mid \mathbf{x},\, \boldsymbol{\phi}\right) \propto \exp\left\{-\frac{\mathbf{a}^T\mathbf{a}}{2\sigma^2_{\mathbf{a}}}\right\} \prod_{i\in\mathcal{L}} \left|\boldsymbol{\Sigma}_i\right|^{-\frac{1}{2}} E_i^{-\left(\frac{N_i}{2}+\alpha_{e,i}\right)}, \tag{20a}$$

with

$$E_j = 2\beta_{e,j} + \mathbf{s}_j^T\mathbf{s}_j - \mathbf{s}_j^T\mathbf{U}_j\, \boldsymbol{\Sigma}_j^{-1}\, \mathbf{U}_j^T\mathbf{s}_j, \tag{20b}$$

and

$$\boldsymbol{\Sigma}_j = \mathbf{U}_j^T\, \mathbf{U}_j + \delta^{-2}_{\mathbf{b}_j}\mathbf{I}_{FQ}, \tag{20c}$$

where $j \in \mathcal{L}$, $\delta_{\mathbf{b}_j}$ is a hyperparameter defined for analytical tractability as $\sigma^2_{\mathbf{b}_j} \triangleq \delta^2_{\mathbf{b}_j}\sigma^2_{e,j}$. In eqn. (20), it is understood that $\mathbf{s}_i$ and $\mathbf{U}_i$ are functions of the parameters $\mathbf{a}$ and the observed data $\mathbf{x}$. The maximum marginal *a posteriori* (MMAP) estimate is found by solving $\hat{\mathbf{a}}_{\text{MMAP}} = \arg\max_{\mathbf{a}} p\left(\mathbf{a} \mid \mathbf{x},\, \boldsymbol{\phi}\right)$.

D. Channel estimation using the Gibbs sampler

In practice, $\hat{\mathbf{a}}_{\text{MMAP}}$ is difficult to find as the *a posteriori* pdf is multi-modal and subject to rapid parameter variation. Instead, MCMC methods can be utilised to sample from the joint pdf of the channel and source parameters as well as the error residual. Gibbs sampling [31], [33]–[35] is a MCMC method that proceeds by iteratively drawing random variates from conditional densities in order to sample from their joint pdf. Independent of the initial distribution, the probabilities of the chain are guaranteed to converge to the invariant distribution, i.e., the joint pdf, after a sufficiently long burn-in period. A minimum mean-square

error (MMSE) estimate of the channel parameters is then obtained through marginalisation of the nuisance parameters by computing the expected value of only the variates of interest.

To sample from the joint pdf of the source coefficients in block i, $\mathbf{b}_i$, the channel coefficients, $\mathbf{a}$, the source and channel hyperparameters, $\sigma^2_{\mathbf{b}_i}$ and $\sigma^2_{\mathbf{a}}$, and the error residual variance, $\sigma^2_{e,i}$, in M runs, the Gibbs sampler iterates for $j \in \mathcal{M}$ through

$$
\begin{aligned}
\mathbf{a}^{(j+1)} &\leftarrow p\left(\mathbf{a} \mid \mathbf{b}^{(j)}, \boldsymbol{\sigma}_e^{(j)}, \boldsymbol{\phi}^{(j)}\right) \\
\mathbf{b}_i^{(j+1)} &\leftarrow p\left(\mathbf{b} \mid \mathbf{a}^{(j+1)}, \boldsymbol{\sigma}_e^{(j)}, \boldsymbol{\phi}^{(j)}\right) \\
\left(\sigma^2_{e,i}\right)^{(j+1)} &\leftarrow p\left(\sigma^2_{e,i} \,\middle|\, \mathbf{a}^{(j+1)}, \mathbf{b}^{(j+1)}, \boldsymbol{\sigma}_{e_{-\sigma^2_{e,j}}}, \boldsymbol{\phi}^{(j)}\right) \\
\left(\sigma^2_{\mathbf{a}}\right)^{(j+1)} &\leftarrow p\left(\sigma^2_{\mathbf{a}} \,\middle|\, \mathbf{a}^{(j+1)}, \mathbf{b}^{(j+1)}, \boldsymbol{\sigma}_e^{(j+1)}, \boldsymbol{\phi}^{(j)}_{-\sigma^{2(j)}_{\mathbf{a}}}\right) \\
\left(\sigma^2_{\mathbf{b}_i}\right)^{(j+1)} &\leftarrow p\left(\sigma^2_{\mathbf{b}_i} \,\middle|\, \mathbf{a}^{(j+1)}, \mathbf{b}_i^{(j+1)}, \boldsymbol{\sigma}_e^{(j+1)}, \boldsymbol{\phi}^{(j)}_{-\sigma^{2(j)}_{\mathbf{b}_i}}\right),
\end{aligned}
$$

where $\boldsymbol{\phi}_{-\boldsymbol{\alpha}}$ denotes $\boldsymbol{\phi}$ with element $\boldsymbol{\alpha}$ removed. The initial distribution $\left\{\mathbf{a}^{(0)}, \mathbf{b}^{(0)}, \boldsymbol{\sigma}_e^{(0)}, \boldsymbol{\phi}^{(0)}\right\}$ is determined randomly or deterministically. The conditionals are derived in Appendix B.

VI. Experimental Results

The results presented in this paper aim to demonstrate the performance of the Bayesian inference for the proposed models for both simulated and real data. As shown in section §III-D, the proposed time-varying all-pole filter shows promise as a channel model; future research will investigate the construction of a complete fullband signal model and evaluate the algorithm for real AIRs. Therefore, the results presented are in a restricted frequency range and for a simplified acoustic channel: fullband signal enhancement could be achieved using the subband method mentioned in section §III-E. The acoustic channel is based on the frequency response of an acoustic gramophone horn, as discussed in [5]. The simulated data is chosen to reflect the statistical nature of speech.

A. Channel Model

In each of the experiments, the acoustic channel is based on perturbations of an actual acoustic gramophone horn response up to a frequency of 1225 Hz [5]. This range matches

that of the investigations in section §III-D. The magnitude frequency response of the original time-invariant channel has four resonant modes which introduces a reasonable and noticeable amount of acoustic distortion into a signal passed through the filter. A time-varying response is obtained by perturbing each of the original channel poles in a circle of small radius. Despite there being a highly nonlinear relationship between the poles and filter parameters, it is possible to accurately model the parameter variation using the sinusoidal basis set:

$$\{g_\ell(n)\} = \{1,\, \sin(2\pi t),\, \cos(2\pi t),\, \sin(2.5\pi t),\, \cos(2.5\pi t)\}$$

[Fig. 7 about here.]

The variability of the channel is shown as the grey lines in Fig. 7. Here, the magnitude frequency response of the channel is plotted at each time instance, assuming the parameters represent an equivalent LTI system. The frequency response of the original unperturbed channel corresponds to the black line; the actual pole variations are shown in Fig. 8b.

B. Single-block TVAR source model

[Fig. 8 about here.]

The first experiment considers globally modelling the source using a single-block TVAR. A synthetic 4^{th}-order TVAR process is used as the input to the 8^{th}-order channel. The source is generated with time-varying parameters that reflect the pole variations of real speech. The parameter variations are chosen to give the least-squares estimate (LSE) approximations of the two leftmost pole trajectories shown in Fig. 3b: these trajectories are reproduced in Fig. 8a to reiterate this. The procedure for determining this approximation is outlined in section §II; thus, the basis set used for the source corresponds to the Fourier set $\{f_k(n)\} = \{\sin(n\omega_0 t),\, \cos(n\omega_0 n)\}_{n=0}^{2}$ with fundamental frequency $\omega_0 = 2\pi\frac{5}{9}$ rad/sec. The total number of source samples used is $N = 2,000$, and is chosen to give sufficient data that the channel estimates have low variance. In practice, of course, this is 4 times the number of samples in Fig. 3b. With regards to eqn. (4), $L = 1$, $T_1 = 4$ and $T_2 = N$, where T_i are the changepoints, i.e., T_1 is the index of the first sample in the block and T_2 is the index of the last sample in the block.

The Gibbs sampler is executed for 5000 iterations with a burn-in period of 500 (10%) samples, although the estimates tend to converge within a few hundred samples. A Monte Carlo experiment with 100 runs is executed to ensure that the performance is consistent and not dependent on the excitation sequence used in the synthetic source. The averaged estimated pole trajectories are shown in Fig. 8. Any individual run gives very similar results to the averaged performance: source and channel pole estimates (grey dots) are relatively close to the actual trajectories (black dots).

Although the channel is identified with reasonable accuracy in the case shown here, in other (unshown) single-block simulations, the MCMC estimates do in fact indicate possible source-channel ambiguities. The multi-block case is more robust to this problem, as discussed in section §IV.

C. Block-based TVAR approach

The single-block TVAR model will not adequately capture the full time-varying nature of a real speech signal and therefore, as discussed in section §II-D, a block-based model is more flexible. To demonstrate the algorithm in this case, the source model in section §VI-B is modified into a multi-block-based time-varying AR model, where the pole variation in each block is smooth, but abrupt change in variation occurs for pole positions between blocks. There are 4 blocks, each 2000 samples long, and the model order in each block is again 4. The pole variations for the source in each block are shown in Figs. 9a and 9b. The source basis functions and settings for the Gibbs sampler are as described in section §VI-B. As can be seen from Figs. 9a, 9b, and 9c, the algorithm is able to accurately detect the pole trajectories. The estimated source, $s_{\text{MMSE}}(n)$ is obtained by solving eqn. (7) with the given channel estimate. An error signal is defined as:

$$\epsilon_{\text{MMSE}}(n) = \left(s_{\text{MMSE}}(n) - s(n)\right)^2 \tag{21}$$

Typical signals are shown in Fig. 10, and a typical histogram of one of the channel parameter samples is shown in Fig. 9d.

[Fig. 9 about here.]

[Fig. 10 about here.]

D. Identifiability issues for real speech signals

As expected, the results for synthetic data generated according to the proposed models demonstrate the estimation algorithm works well, and good enhancement is possible. This is subject to the discussion in section §IV in which it is seen that sufficient pole movement near the unit circle is required for identifiability. With regards to arbitrarily chosen real speech data, often the source pole movement is not sufficient for identifiability. Therefore, it is necessary to identify blocks of observed data for which there is enough pole movement for good parameter estimation. But what is 'sufficient movement'?

The pole variations in Fig. 9 are such that, on average, there is sufficient energy at different spectral regions for the channel to be estimated correctly. In the following, the simulation in section §VI-C is repeated, with the same parameters, except the pole trajectories in Fig. 9 are shortened by a scalar factor ρ. When $\rho = 1$, the poles follow the full variation in Figs. 9a and 9b; when $\rho = 0$, the poles are fixed and stationary at the initial pole position. The basis functions used in section §VI-A are still appropriate for modelling these variations. Defining the log normalised estimation error as $\hat{\epsilon}_{\text{MMSE}}(n) = 20 \log_{10} \epsilon_{\text{MMSE}}(n)/s^2(n)$ from eqn. (21), Fig. 11 shows $\hat{\epsilon}_{\text{MMSE}}(n)$ as a function of pole variability: a source with larger variability in pole movements leads to improved signal enhancement. In particular, note the block-stationary model does not provide enough 'spectral excitation' for good channel identification compared to the TVAR model.

[Fig. 11 about here.]

E. Results for real speech

Fig. 12 shows results for the case when real speech is filtered through the channel. The 'true source poles' (black dots) are estimated from the known clean speech for comparison – model-order 6. Again, the basis functions and Gibbs sampler settings are as in section §VI-B. The variability of the poles in Fig. 12, estimated from arbitrary segments of speech, is significant. As predicted in sections §IV and §VI-D, they are thus more difficult to estimate,

and the channel estimates, although in the right regions, are considerably off compared to the simulated examples. There is, however, still a 2.2 dB reduction in signal error and thus some speech enhancement.

[Fig. 12 about here.]

An open question remains about the choice of model order for the source signal. Whereas for BSAR speech models, the model order is generally greater than, say, 15, in these experiments the source is modelled as a low-order TVAR process where, say, 5 basis functions are needed to model each parameter. Thus, for a 6-th order model, 30 *parameters* must be estimated. Since the TVAR process is more flexible than a BSAR model, can lower model orders be used?

VII. Conclusions

Blind dereverberation of speech from a moving speaker is a challenging problem that has a number of practical applications. A previous approach to single-channel blind dereverberation [5] focusing on stationary speakers assumed a locally-stationary source signal and uses the time-invariance of the channel to resolve estimation ambiguities. This paper provides a novel contribution towards approaching single-channel blind dereverberation from a moving speaker by utilising a more general and flexible block-based TVAR process to model the speech signal, and a LTV all-pole filter for the acoustic channel. Simulations show that the channel estimates are more accurate when the multi-block model is used over the single-block case, and it is argued the multi-block model provides the necessary flexibility for modelling long segments of speech.

A Bayesian inference algorithm is developed to estimate the system parameters. As expected, simulated results show that parameter estimates are good when the data fits the model. Substantial discussion is given justifying the models used for real data. Further work includes: 1) dealing with speech segments in which the spectral excitation is weak; 2) further model validation, and algorithmic testing on data obtained in a realistic acoustic environments; 3) developing an algorithm that does not implicitly rely on inverse filtering of the channel which would fail for ill-conditioned channels; 4) utilising subband models; 4) dealing with

non-minimum phase channels that are the norm for real acoustic environments.

APPENDIX A
POSTERIOR PDF OF CHANNEL PARAMETERS

Inserting the priors in section §V-B and the likelihood from eqn. (17) into the joint pdf in eqn. (18) gives:

$$\begin{aligned} p\left(\mathbf{a},\mathbf{b},\boldsymbol{\sigma}_e \mid \mathbf{x},\boldsymbol{\phi}\right) &\propto \exp\left\{-\frac{\mathbf{a}^T\mathbf{a}}{2\sigma_{\mathbf{a}}^2}\right\} \\ &\times \prod_{i=1}^{L} \frac{1}{\left(\sigma_{e,i}^2\right)^{\frac{N_i}{2}}} \exp\left\{-\frac{1}{2\sigma_{e,i}^2}\left\|\mathbf{s}_i+\mathbf{U}_i\mathbf{b}_i\right\|^2\right\} \\ &\times \frac{1}{\left(\sigma_{e,i}^2\right)^{FQ}} \exp\left\{-\frac{\mathbf{b}_i^T\mathbf{b}_i}{2\delta_{\mathbf{b}_i}^2\,\sigma_{e,i}^2}\right\} \frac{1}{\left(\sigma_{e,i}^2\right)^{\alpha_{e,i}+1}} \exp\left\{-\frac{\beta_{e,i}}{\sigma_{e,i}^2}\right\}, \end{aligned} \tag{22}$$

where for analytical tractability, set $\sigma_{\mathbf{b}_k}^2 = \delta_{\mathbf{b}_k}^2\,\sigma_{e,j}^2$, where $k \in \mathcal{L}$, $\delta_{\mathbf{b}_k}^2$ is a hyperparameter, and also where terms that involve hyperparameters have been ignored since they are assumed known. To obtain $p\left(\mathbf{a},\mathbf{b}_{-\mathbf{b}_k},\boldsymbol{\sigma}_e \mid \mathbf{x},\boldsymbol{\phi}\right)$, marginalise $\mathbf{b}_k$:

$$\begin{aligned} p\left(\mathbf{a},\mathbf{b}_{-\mathbf{b}_k},\boldsymbol{\sigma}_e \mid \mathbf{x},\boldsymbol{\phi}\right) &\propto \exp\left\{-\frac{\mathbf{a}^T\mathbf{a}}{2\sigma_{\mathbf{a}}^2}\right\} \prod_{\ell=1}^{L} \frac{1}{\left(\sigma_{e,\ell}^2\right)^{R_\ell}} \\ &\times \prod_{j\neq i=1}^{L} \exp\left\{\frac{-1}{2\sigma_{e,i}^2}\left(\left\|\mathbf{s}_i+\mathbf{U}_i\mathbf{b}_i\right\|^2+\frac{\mathbf{b}_i^T\mathbf{b}_i}{2\delta_{\mathbf{b}_i}^2}+2\beta_{e,i}\right)\right\} \\ &\times \underbrace{\int_{-\infty}^{\infty} \exp\left\{\frac{-1}{2\sigma_{e,j}^2}\left(\left\|\mathbf{s}_j+\mathbf{U}_j\mathbf{b}_j\right\|^2+\frac{\mathbf{b}_j^T\mathbf{b}_j}{2\delta_{\mathbf{b}_j}^2}+2\beta_{e,j}\right)\right\} d\mathbf{b}_j}_{\Phi_j}, \end{aligned} \tag{23}$$

with $R_k = \frac{N_k+FQ+2(\alpha_{e,k}+1)}{2}$. Writing the integrand of Φ_k as:

$$\begin{aligned} I_k = \exp\Bigg\{ -\frac{1}{2\sigma_{e,k}^2}\Bigg\{ & 2\beta_{e,k}+\mathbf{s}_k^T\mathbf{s}_k+2\mathbf{s}_k^T\mathbf{U}_k\,\mathbf{b}_k+ \\ & \times \mathbf{b}_k^T\left(\mathbf{U}_k^T\,\mathbf{U}_k+\delta_{\mathbf{b}_k}^{-2}\mathbf{I}_{FQ}\right)\mathbf{b}_k\Bigg\}\Bigg\}. \end{aligned}$$

Comparing the integral Φ_k with the standard Gaussian identity,

$$\int_{\mathbb{R}^P} \exp\left\{-\frac{1}{2}\left[\alpha + 2\boldsymbol{\beta}^T\mathbf{y} + \mathbf{y}^T\boldsymbol{\Gamma}\mathbf{y}\right]\right\} d\mathbf{y} = \frac{(2\pi)^{\frac{P}{2}}}{|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left[\alpha - \boldsymbol{\beta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}\right]\right\},$$

where E_k and $\boldsymbol{\Sigma}_k$ are defined in section §V-C. Thus:

$$\Phi_k = \frac{(2\pi\sigma_{e,k}^2)^{\frac{FQ}{2}}}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{-\frac{E_k}{2\sigma_{e,k}^2}\right\}.$$

Hence, eqn. (23) simplifies, and repeating over all k:

$$p\left(\mathbf{a}, \boldsymbol{\sigma}_e \mid \mathbf{x}, \boldsymbol{\phi}\right) \propto \exp\left\{-\frac{\mathbf{a}^T\mathbf{a}}{2\sigma_{\mathbf{a}}^2}\right\} \times \prod_{i=1}^{L} \frac{1}{|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \underbrace{\frac{1}{\left(\sigma_{e,i}^2\right)^{R_i - \frac{FQ}{2}}} \exp\left\{-\frac{E_i}{2\sigma_{e,i}^2}\right\}}_{\Psi_i}. \tag{24}$$

The conditional pdf of the channel parameters, $p\left(\mathbf{a} \mid \mathbf{x}, \boldsymbol{\phi}\right)$, is found by marginalising the error residual variance, $\sigma_{e,i}^2$:

$$p\left(\mathbf{a} \mid \mathbf{x}, \boldsymbol{\phi}\right) \propto \int_0^\infty \cdots \int_0^\infty p\left(\mathbf{a}, \boldsymbol{\sigma}_e \mid \mathbf{x}, \boldsymbol{\phi}\right) \, d\sigma_{e,L}^2 \cdots \sigma_{e,1}^2. \tag{25}$$

The integrand, Ψ_i, from eqn. (24) is solved using the identity:

$$\int_0^\infty \frac{1}{\left(\sigma^2\right)^{(\beta+1)}} \exp\left\{-\frac{\alpha}{\sigma^2}\right\} d\sigma^2 = \frac{\Gamma(\beta)}{\alpha^\beta}.$$

Since the $\{\sigma_{e,i}^2\}$'s are independent, eqn. (20a) thus follows.

APPENDIX B
GIBBS SAMPLER – CONDITIONAL PDFS

According to Bayes's theorem, the conditional pdfs are:

$$p\left(\mathbf{a} \mid \boldsymbol{\theta}_{-\mathbf{a}}\right) \propto p\left(\mathbf{x} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e\right) p\left(\mathbf{a} \mid \sigma_{\mathbf{a}}^2\right) \tag{26a}$$

$$p\left(\mathbf{b}_i \mid \boldsymbol{\theta}_{-\mathbf{b}_i}\right) \propto p\left(\mathbf{x} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e\right) p\left(\mathbf{b}_i \mid \sigma_{\mathbf{b}_i}^2\right) \tag{26b}$$

$$p\left(\sigma_{\mathbf{a}}^2 \mid \boldsymbol{\theta}_{-\sigma_{\mathbf{a}}^2}\right) \propto p\left(\mathbf{a} \mid \sigma_{\mathbf{a}}^2\right) p\left(\sigma_{\mathbf{a}}^2 \mid \alpha_{\mathbf{a}}, \beta_{\mathbf{a}}\right) \tag{26c}$$

$$p\left(\sigma^2_{\mathbf{b}_i} \,\middle|\, \boldsymbol{\theta}_{-\sigma^2_{\mathbf{b}_i}}\right) \propto p\left(\mathbf{b}_i \,\middle|\, \sigma^2_{\mathbf{b}_i}\right) p\left(\sigma^2_{\mathbf{b}_i} \,\middle|\, \alpha_{\mathbf{b}_i}, \beta_{\mathbf{b}_i}\right) \tag{26d}$$

$$p\left(\sigma^2_{e,i} \,\middle|\, \boldsymbol{\theta}_{-\sigma^2_{e,i}}\right) \propto p\left(\mathbf{x} \,\middle|\, \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e\right) p\left(\sigma^2_{e,i} \,\middle|\, \alpha_{e,i},\, \beta_{e,i}\right) \tag{26e}$$

where $\boldsymbol{\theta} = \{\mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e, \boldsymbol{\phi}\}$. The likelihood function is found in eqn. (17), and the priors are defined in section §V-B.

A. Channel coefficients

Recall eqn. (9b), $\mathbf{e}_i = \mathbf{B}_i\,\hat{\mathbf{s}}_i$, such that the likelihood is given by eqn. (16). It is desired to write $\hat{\mathbf{s}}_i$ as a linear function of $\mathbf{a}$. Consider writing the model in eqn. (7) for the values needed to define $\hat{\mathbf{s}}_i$ in eqn. (9b): $n \in \{T_i - Q,\, \ldots,\, T_{i+1} - 1\}$. Let $\hat{N}_i = N_i + Q$, and the $\hat{N}_i \times 1$ vector

$$\mathbf{x}_{i,p} = \begin{bmatrix} x(T_i - Q - p) & \cdots & x(T_{i+1} - 1 - p) \end{bmatrix}^T.$$

Then $\hat{\mathbf{x}}_i = \mathbf{x}_{i,0}$ is a $\hat{N}_i \times 1$ vector and $\mathbf{X}_{i,p} = \operatorname{diag}\left[\mathbf{x}_{i,p}\right]$ is a $\hat{N}_i \times \hat{N}_i$ diagonal matrix; $\mathbf{G}_{i,p}$ is a $N_i \times G$ matrix whose columns are the G basis functions evaluated between $n = \{T_i - Q - p,\, \ldots,\, T_{i+1} - 1 - p\}$, such that the (j,k)-th element of $\mathbf{G}_{i,p}$ is $[\mathbf{G}_{i,p}]_{jk} = g_k(j + T_i + Q - q)$. Hence, it follows that $\mathbf{V}_i$ is the $N_i \times GP$ matrix $\mathbf{V}_i = \begin{bmatrix} \mathbf{V}_{i,1} & \cdots & \mathbf{V}_{i,P} \end{bmatrix}$ where $\mathbf{V}_{i,p} = \mathbf{X}_{i,p}\,\mathbf{G}_{i,p}$, and $\mathbf{a} = \begin{bmatrix} \mathbf{a}_1^T & \ldots & \mathbf{a}_P^T \end{bmatrix}^T$ is a $GP \times 1$ vector, where $[\mathbf{a}_p]_k = a_{p\ell}$. This is equivalent to writing:

$$\hat{\mathbf{s}}_i = \hat{\mathbf{x}}_i + \mathbf{V}_i \mathbf{a}.$$

Substituting into eqn. (16) gives:

$$p_{\mathbf{X}}\left(\mathbf{x} \,\middle|\, \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}_e\right) \propto \left[\prod_{i=1}^{L} \frac{1}{\sigma_{e,i}^{N_i}} \exp\left\{-\frac{\left\|\mathbf{B}_i\,\hat{\mathbf{x}}_i + \mathbf{B}_i\mathbf{V}_i\mathbf{a}\right\|^2}{2\sigma^2_{e,i}}\right\}\right]_{\mathbf{s}=\mathbf{A}\mathbf{x}}.$$

Defining $\mathbf{V}_{\mathbf{b}_i} = \mathbf{B}_i\mathbf{V}_i$ and $\mathbf{x}_{\mathbf{b}_i} = \mathbf{B}_i\hat{\mathbf{x}}_i$, inserting $p\left(\mathbf{a} \,\middle|\, \sigma^2_{\mathbf{a}}\right)$ and this likelihood into eqn. (26a), it follows the conditional pdf of the channel coefficients is multivariate Gaussian, $p\left(\mathbf{a} \,\middle|\, \mathbf{x}, \mathbf{b}, \boldsymbol{\sigma}_e, \boldsymbol{\phi}\right) = \mathcal{N}\left(\mathbf{a} \,\middle|\, \boldsymbol{\mu}_{\mathbf{a}},\, \boldsymbol{\Gamma}_{\mathbf{a}}\right)$, with inverse covariance

$$\boldsymbol{\Gamma}_{\mathbf{a}}^{-1} = \frac{\mathbf{I}_{GP}}{\sigma^2_{\mathbf{a}}} + \sum_{i=1}^{L} \frac{1}{\sigma^2_{e,i}} \mathbf{V}_{\mathbf{b}_i}^T \mathbf{V}_{\mathbf{b}_i} \tag{27}$$

and mean
$$\boldsymbol{\mu}_{\mathbf{a}} = -\boldsymbol{\Gamma}_{\mathbf{a}} \sum_{i=1}^{L} \frac{1}{\sigma_{e,i}^2} \mathbf{V}_{\mathbf{b}_i}^T \, \mathbf{x}_{\mathbf{b}_i}, \tag{28}$$

Note that the vector $\mathbf{x}_{\mathbf{b}_i}$ can be calculated efficiently by writing $\mathbf{x}_{\mathbf{b}_i} = \mathbf{B}_i \hat{\mathbf{x}}_i = \tilde{\mathbf{x}}_{i,0} + \mathbf{W}_i \, \mathbf{b}_i$, where

$$\tilde{\mathbf{x}}_{i,q} = \begin{bmatrix} x(T_i - q) & \cdots & x(T_{i+1} - 1 - q) \end{bmatrix}^T,$$

and $\mathbf{W}_i$ is the $N_i \times FQ$ matrix $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{i,1} & \cdots & \mathbf{W}_{i,Q} \end{bmatrix}$ where $\mathbf{W}_{i,q} = \tilde{\mathbf{X}}_{i,q} \, \mathbf{F}_{i,q}$ with $\tilde{\mathbf{X}}_{i,q} = \text{diag}\,[\tilde{\mathbf{x}}_{i,q}]$. Similarly, observe that each column of the matrix $\mathbf{V}_{i,\mathbf{b}} = \mathbf{B}_i \mathbf{V}_i$ can also be written in a similar fashion; thus, defining $\mathbf{v}_{i,r} = [\mathbf{V}_i]_r$ as being the r-th column of $\mathbf{V}_i$, then $\mathbf{B}_i \, \mathbf{v}_{i,r} \equiv \tilde{\mathbf{v}}_{i,r} + \mathbf{W}_{i,r} \, \mathbf{b}_i$ using similar definitions to above.

B. Source coefficients

Using eqn. (17) and $p\left(\mathbf{b}_i \mid \sigma_{\mathbf{b}_i}^2\right)$, then from eqn. (26b):

$$p\left(\mathbf{b}_i \mid \mathbf{x}, \mathbf{a}, \boldsymbol{\sigma}_e, \, \boldsymbol{\phi}\right) \propto \exp\left\{ -\frac{1}{2} \left[\frac{2}{\sigma_{e,i}^2} \mathbf{s}_i^T \, \mathbf{U}_i \, \mathbf{b}_i + \mathbf{b}_i^T \left(\frac{1}{\sigma_{e,i}^2} \mathbf{U}_i^T \mathbf{U}_i + \frac{\mathbf{I}_{FQ}}{\sigma_{\mathbf{a}}^2} \right) \mathbf{b}_i \right] \right\}_{\mathbf{s}=\mathbf{A}\mathbf{x}}.$$

Define $\mathbf{x}_{\mathbf{a},i}$ as the vector $\mathbf{s}_i$ with $\mathbf{s} = \mathbf{A}\mathbf{x}$, and similarly $\mathbf{Y}_i$ as the matrix $\mathbf{U}_i$ with samples $s(n)$ replaced by $\mathbf{s} = \mathbf{A}\mathbf{x}$. Then, the conditional pdf of the source parameters is also Gaussian, $p\left(\mathbf{b}_i \mid \mathbf{x}, \mathbf{a}, \boldsymbol{\sigma}_e, \boldsymbol{\phi}\right) = \mathcal{N}\left(\mathbf{b}_i \,\middle|\, \boldsymbol{\mu}_{\mathbf{b}_i}, \, \boldsymbol{\Gamma}_{\mathbf{b}_i}\right)$, with:

inverse covariance
$$\boldsymbol{\Gamma}_{\mathbf{b}_i}^{-1} = \frac{1}{\sigma_{e,i}^2} \mathbf{Y}_i^T \mathbf{Y}_i + \frac{1}{\sigma_{\mathbf{b}_i}^2} \mathbf{I}_{GF} \tag{29}$$

and mean
$$\boldsymbol{\mu}_{\mathbf{b}_i} = -\frac{1}{\sigma_{e,i}^2} \boldsymbol{\Gamma}_{\mathbf{b}_i} \mathbf{Y}_i^T \, \mathbf{x}_{\mathbf{a},i}, \tag{30}$$

Consider the term $\mathbf{x}_{\mathbf{a},i} = \mathbf{A}_i \, \mathbf{x}$, where the matrix $\mathbf{A}_i \in \mathbb{R}^{N_i \times N_i}$ is defined appropriately. This can be calculated efficiently by writing it in the form $\mathbf{x}_{\mathbf{a},i} = \bar{\mathbf{x}}_0 + \bar{\mathbf{W}} \, \mathbf{a}$.

C. Error residual variance and hyperparameters

Defining $E_i = \|\mathbf{x}_{\mathbf{a},i} + \mathbf{Y}_i \, \mathbf{b}_i\|^2$, and inserting the likelihood and $p\left(\sigma_{e,i}^2 \,\middle|\, \alpha_{e,i}, \, \beta_{e,i}\right)$ into eqn. (26e), it follows the error residual variance has an inverse-Gamma distribution:

$$p\left(\sigma_{e,i}^2 \,\middle|\, \mathbf{x}, \mathbf{a}, \mathbf{b}, \boldsymbol{\phi}\right) \sim \mathcal{IG}\left(\alpha_{e,i} + N_i/2, \; E_i/2 + \beta_{e,i}\right) \tag{31}$$

Similarly, the sampling distribution for the hyperparameters of the source and channel coefficients are found from eqns. (26c) and (26d) to be inverse-Gamma distributions: $p\left(\sigma^2_{\mathbf{a}} \mid \mathbf{x}, \mathbf{a}, \mathbf{b}, \boldsymbol{\phi}\right) \sim \mathcal{IG}\left(PG/2 + \alpha_{\mathbf{a}},\, \mathbf{a}^T\mathbf{a}/2 + \beta_{\mathbf{a}}\right)$, and $p\left(\sigma^2_{\mathbf{b}_i} \,\middle|\, \mathbf{x}, \mathbf{a}, \mathbf{b}_i, \boldsymbol{\phi}\right) \sim \mathcal{IG}\left(QF/2 + \alpha_{\mathbf{b}_i},\, \mathbf{b}_i^T\mathbf{b}_i/2 + \beta_{\mathbf{b}_i}\right)$.

ACKNOWLEDGEMENT

The authors would like to thank Dr Steven Fortune for assistance in measuring the real AIRs, and to Xionghu Zhong for setting up the image-method simulations, both used in section §III-D. The authors would also like to thank the anonymous reviewers for their valuable suggestions and comments.

REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, February 1988.

[2] S. Weiss, G. W. Rice, and R. W. Stewart, "Multichannel equalization in subbands," in *Proc. IEEE Conf. WASPAA*, Mohonk Mountain House, NY, October 1999, pp. 203–206.

[3] M. Daly, J. P. Reilly, and J. Manton, "A Bayesian approach to blind source recovery," in *Asil. Conf. Signals, Syst., Comput.*, Asilomar, Pacific Grove, CA, 2004.

[4] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 80–95, January 2007.

[5] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 476–488, September 2003.

[6] J. J. Rajan and P. J. W. Rayner, "Parameter estimation of time-varying autoregressive models using the Gibbs sampler," *Electr. Letters*, vol. 31, no. 13, pp. 1035–1036, June 1995.

[7] J. J. Rajan, P. J. W. Rayner, and S. J. Godsill, "Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler," *IEE Proc.-Vis. Image Signal Process.*, vol. 144, no. 4, pp. 249–256, August 1997.

[8] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[9] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley, 1987.

[10] E. Bognar and H. Fujisaki, "Analysis, synthesis and perception of the French nasal vowels," in *Proc. IEEE Conf. ICASSP*, vol. 11, April 1986, pp. 1601–1604.

[11] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[12] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model Based Approach*. Berlin, Germany: Springer Verlag, 1998.

[13] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, pp. 899–011, August 1983.

[14] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Proc.*, vol. 5, no. 3, pp. 267–285, May 1978.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing,*, A. v. Oppenheim, Ed. New Jersey: Prentice Hall Signal Processing Series, 1993, vol. 1.

[16] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 173–185, March 2002.

[17] L. A. Liporace, "Linear estimation of nonstationary signals," *J. Acoust. Soc. Amer.*, vol. 58, no. 6, pp. 1288–1295, December 1976.

[18] R. Charbonnier, M. Barlaud, G. Alengrin, and J. Menez, "Results on AR-modelling of nonstationary signals," *Signal Proc.*, vol. 12, no. 2, pp. 143–151, March 1987.

[19] J. J. Rajan and P. J. W. Rayner, "Generalized feature extraction for time-varying autoregressive models," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2498–2507, October 1996.

[20] T. S. Rao, "The fitting of nonstationary time-series models with time-dependent parameters," *J. Royal Stat. Soc. B*, vol. 32, no. 2, pp. 312–322, 1970.

[21] H. Kuttruff, *Room Acoustics*, 4th ed. London, England: Spon Press, 2000.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, April 1979.

[23] J. N. Mourjopoulos and M. A. Paraskevas, "Pole and zero modeling of room transfer functions," *J. Sound Vibr.*, vol. 146, no. 2, pp. 281–302, April 1991.

[24] J. N. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound Vibr.*, vol. 102, no. 2, pp. 217–228, September 1985.

[25] J. R. Hopgood, "A subband modelling approach to the enhancement of speech captured in reverberant acoustic environments: MIMO case," in *Proc. IEEE Conf. WASPAA*, Mohonk, Oct. 2005.

[26] B. D. Radlović, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Trans. SAP*, vol. 8, no. 3, pp. 311–319, May 2000.

[27] L. G. Johansen and P. Rubak, "The excess phase in loudspeaker/room transfer functions: Can it be ignored in equalization tasks?" *J. of the AES (abstracts)*, May 1996, preprint 4181.

[28] B. D. Radlović and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. SAP*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

[29] H. Wang and F. Itakura, "Dereverberation of speech signals based on sub-band envelope estimation," *IEICE Trans. Fund. Elec. Comms. Comp. Sci.*, vol. E74, no. 11, pp. 3576–3583, Nov. 1991.

[30] ——, "Realization of acoustic inverse filtering through multi-microphone sub-band processing," *IEICE Trans. Fund. Elec. Comms. Comp. Sci.*, vol. E75-A, no. 11, pp. 1474–1483, Nov. 1992.

[31] J. J. K. O. Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York: Springer Verlag, 1996.

[32] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. New York: Holden-Day, 1994.

[33] A. E. Gelfand and A. F. M. Smith, "Sampling based approaches to calculating marginal densities," *J. Am. Stat. Assoc.*, vol. 85, pp. 398–409, 1990.

[34] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, pp. 721–741, 1984.

[35] A. Doucet and X. Wang, "Monte carlo methods for signal processing: a review in the statistical signal processing context," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 152–170, November 2005.

LIST OF FIGURES

1 Model based approach to blind dereverberation 35
2 Speech segment; shaded areas are of length 204 msec or 500 samples at sampling frequency of $f_s = 2.45$ kHz. 36
3 Pole and parameter variations from the speech segment in Fig. 2 for model order $Q = 6$ and 8, block length $N = 500$, $L = N$ steps, sampling frequency $f_s = 2.45$ kHz. 37
4 Block-based time-varying model 38
5 Source and sensor locations in experimental set-up; all measurements in millimeters. Source and sensor elevation is 845 mm, room height of 3200 mm. The sensor is moved downwards from its initial position in 2 mm increments. 39
6 Simulated and experimental results for spatio-temporal variation of the poles in all-pole modelling of AIRs; pole trajectories illustrated through colour map from black (starting point) to light grey (ending point). Model order: 16. 40
7 Equivalent frequency response variation of the LTV all-pole channel 41
8 Actual poles (black dots) vs. Gibbs sampler estimates (grey dots) using 5000 Gibbs sampler iterations, burn-in period of 500 samples, and 100 runs for Monte Carlo simulation. 42
9 Pole trajectories in block-based simulation. Actual poles indicated by black dots, blind estimates by grey dots. 43
10 Observed, source, estimated, and error signals. Vertical line denotes the change-point position. 44
11 Estimation error as function of pole variability, ρ. 45
12 Pole trajectories for real speech signal. Clean speech poles indicated by black dots, blind estimates by grey dots. 46

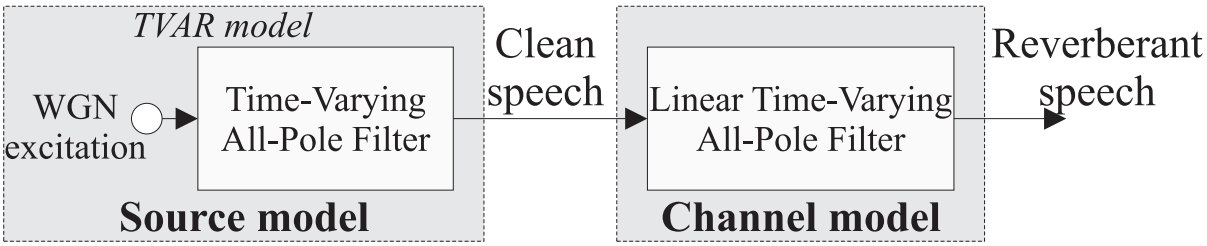


Fig. 1: *Model based approach to blind dereverberation*

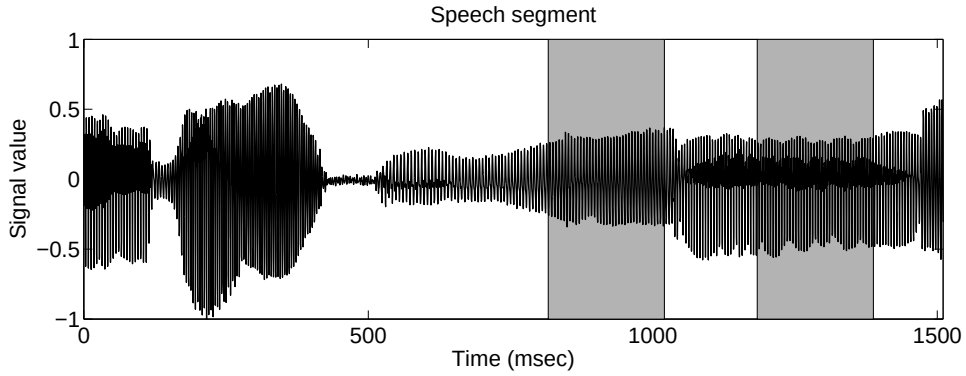
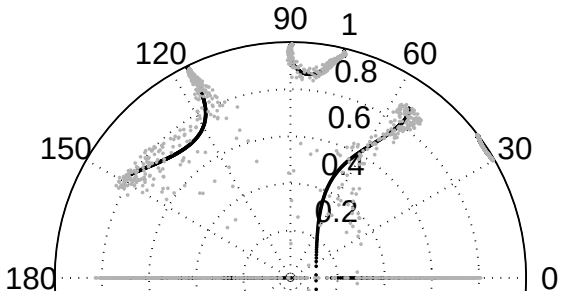
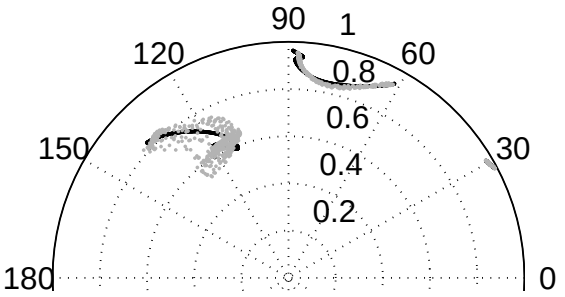


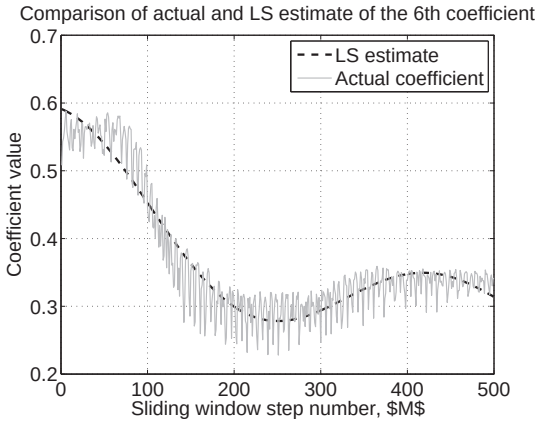
Fig. 2: *Speech segment; shaded areas are of length* 204 *msec or* 500 *samples at sampling frequency of* $f_s = 2.45$ *kHz.*



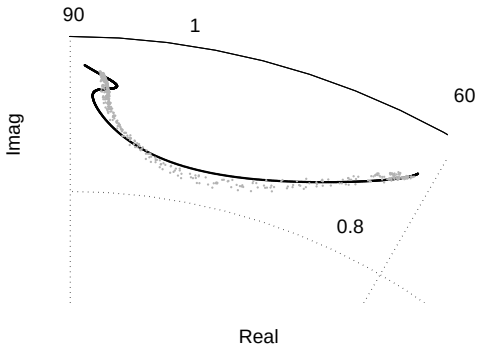
(a) *Birth and death of poles for true poles (grey dots) and LSE (black dots) for the left shaded area in Fig. 2. Model order:* $Q = 8$.



(b) *True poles (grey dots) and LSE (black dots) for speech segment in right shaded area of Fig. 2. Model order:* $Q = 6$.



(c) *Smooth pole variation (Fig. 3b) corresponds to relatively smooth parameter variation.*



(d) *Close-up of Fig. 3b showing LSE (black dots) matching true poles (grey dots).*

Fig. 3: *Pole and parameter variations from the speech segment in Fig. 2 for model order* $Q = 6$ *and* 8*, block length* $N = 500$*,* $L = N$ *steps, sampling frequency* $f_s = 2.45$ *kHz.*

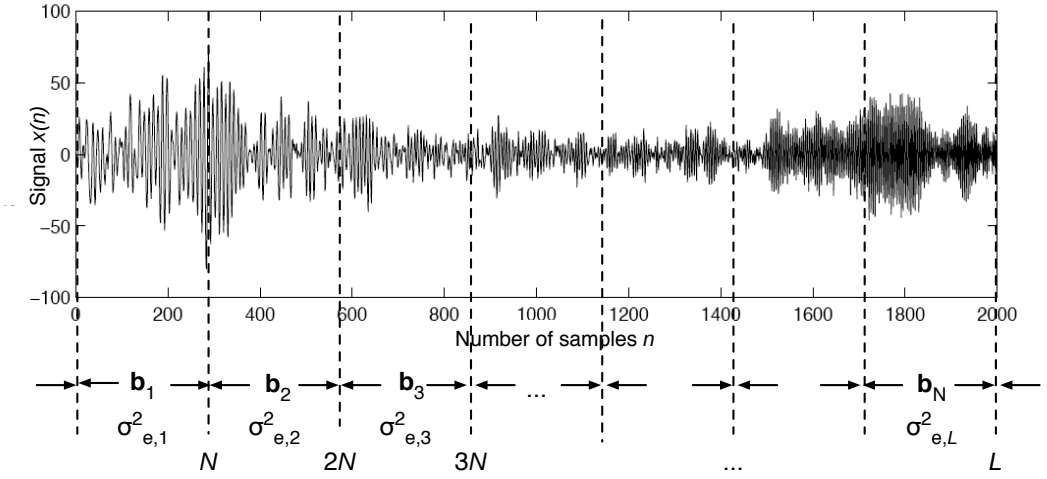


Fig. 4: *Block-based time-varying model*

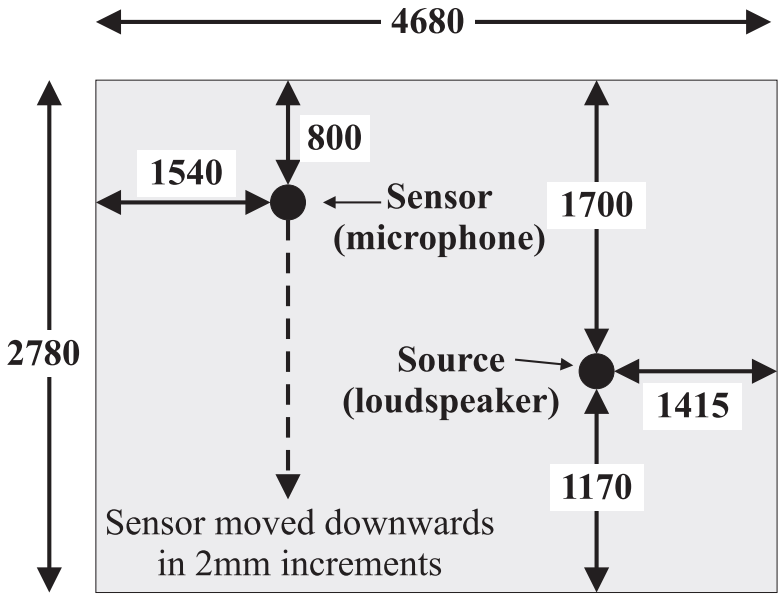
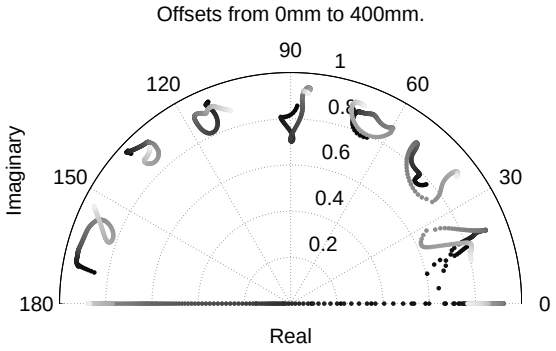
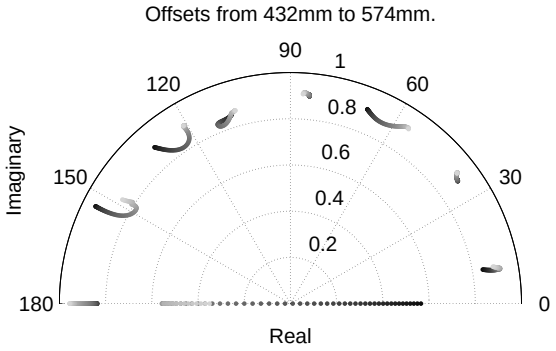


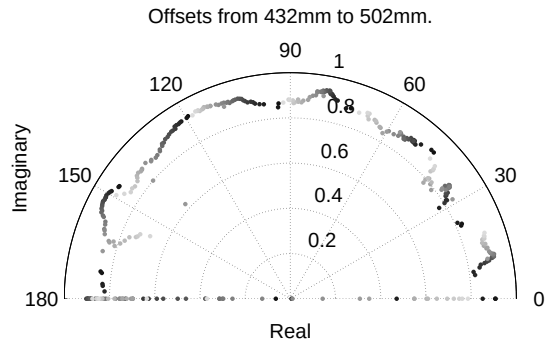
Fig. 5: *Source and sensor locations in experimental set-up; all measurements in millimeters. Source and sensor elevation is* 845 *mm, room height of* 3200 *mm. The sensor is moved downwards from its initial position in* 2 *mm increments.*



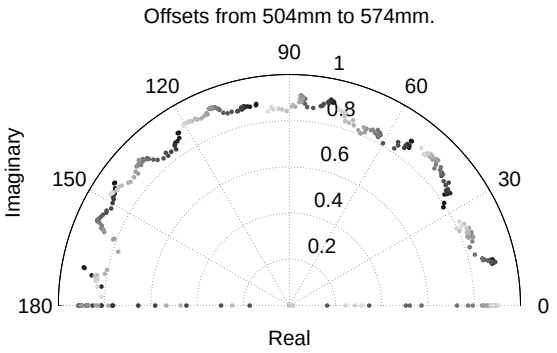
(a) *Simulated: offsets* 0 *to* 400*mm*



(b) *Simulated: offsets* 432 *to* 574*mm*



(c) *Measured: offsets* 432 *to* 502*mm (corresponding to variation observed at mic.* 7*)*



(d) *Measured: offsets* 504 *to* 574*mm (corresponding to variation observed at mic.* 8*)*

Fig. 6: *Simulated and experimental results for spatio-temporal variation of the poles in all-pole modelling of AIRs; pole trajectories illustrated through colour map from black (starting point) to light grey (ending point). Model order:* 16*.*

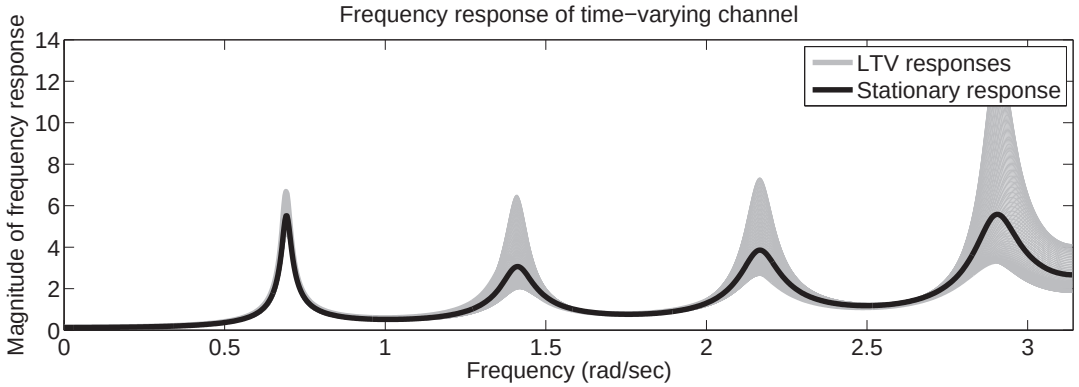
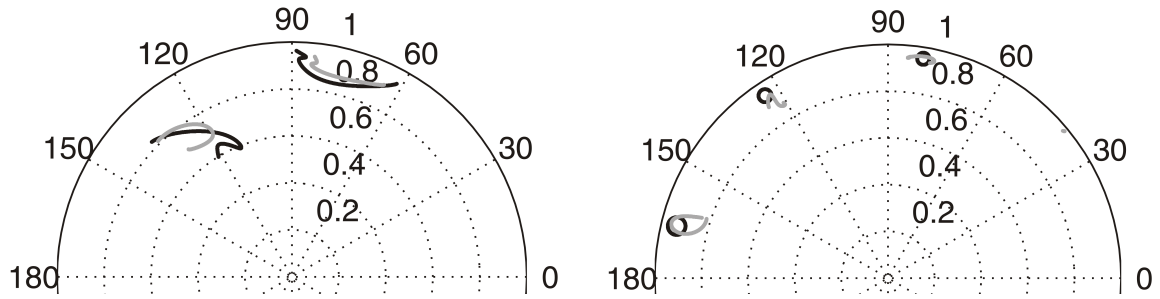


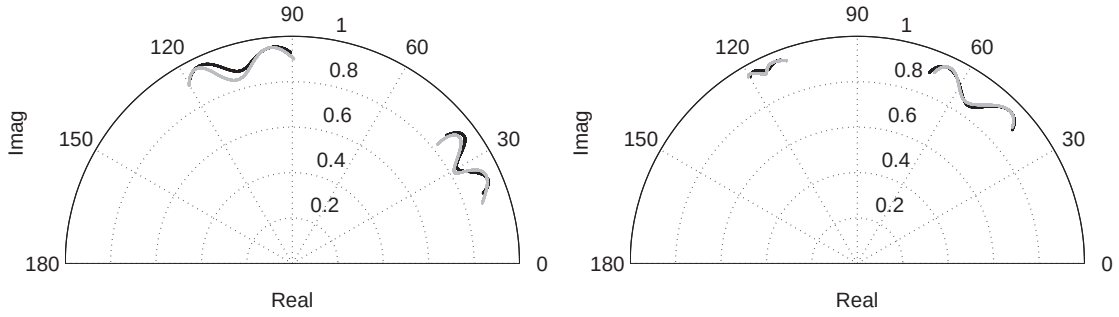
Fig. 7: *Equivalent frequency response variation of the LTV all-pole channel*

.

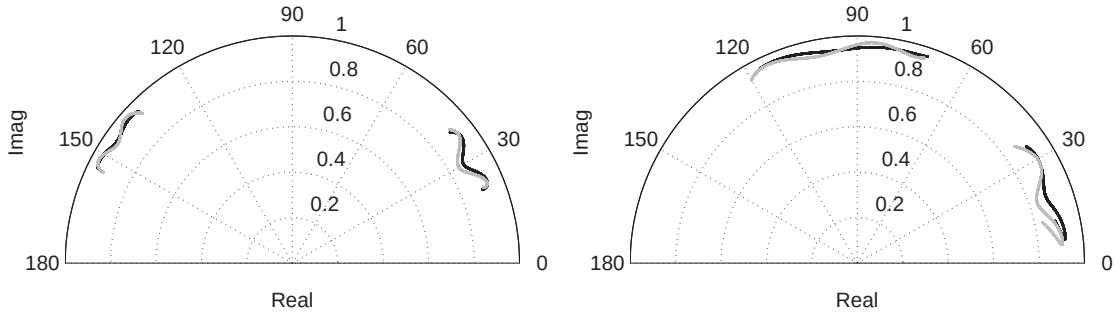


(a) *Time-varying source* **(b)** *Time-varying channel*

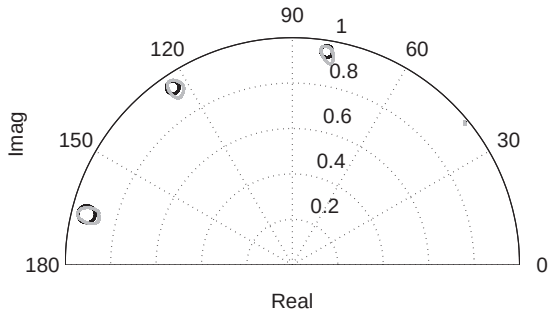
Fig. 8: *Actual poles (black dots) vs. Gibbs sampler estimates (grey dots) using* 5000 *Gibbs sampler iterations, burn-in period of* 500 *samples, and* 100 *runs for Monte Carlo simulation.*



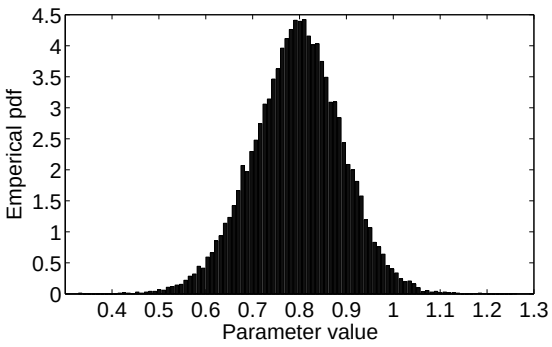
(a) Source poles: blocks 1 (left) and 2 (right)



(b) Source poles: block 3 (left) and 4 (right)



(c) Channel poles



(d) Channel histogram

Fig. 9: *Pole trajectories in block-based simulation. Actual poles indicated by black dots, blind estimates by grey dots.*

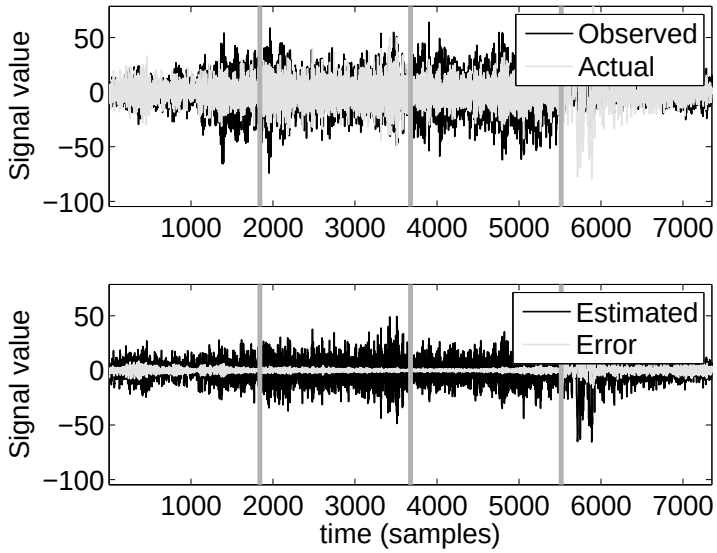


Fig. 10: *Observed, source, estimated, and error signals. Vertical line denotes the changepoint position.*

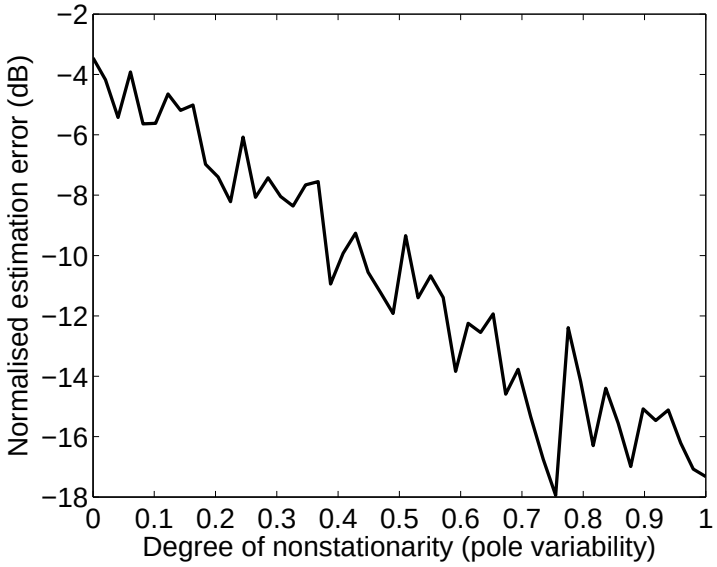
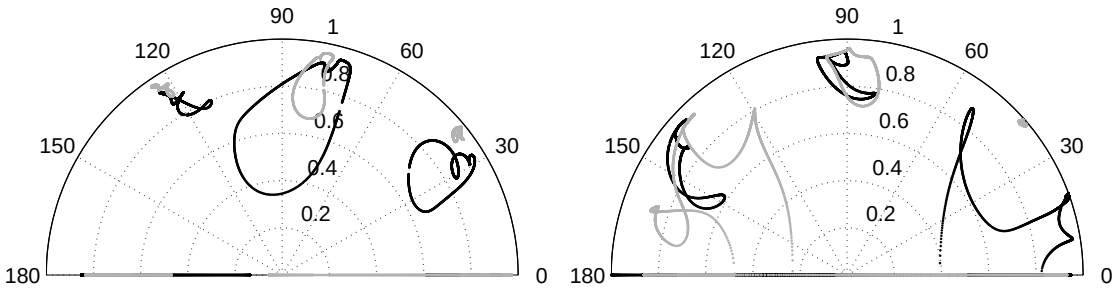
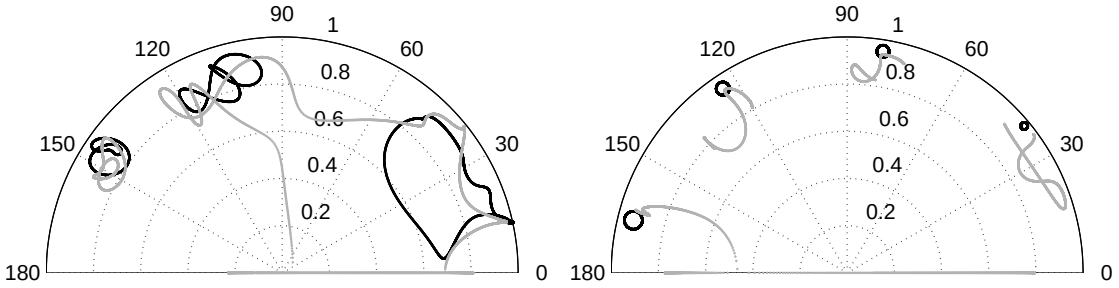


Fig. 11: *Estimation error as function of pole variability, ρ.*



(a) Source poles: blocks 1 (left) and 2 (right)



(b) Source poles: block 3

(c) Channel poles

Fig. 12: *Pole trajectories for real speech signal. Clean speech poles indicated by black dots, blind estimates by grey dots.*

LIST OF TABLES

I The advantages (+) and disadvantages (-) of stationarity and nonstationarity on a local and global level 48

	Stationarity	**Nonstationarity**
Local	- not modelling parameter variation + simpler model	+ model smooth parameter variation - cannot model abrupt changes
Global	- discontinuities at block boundaries + simpler model	+ discontinuities of boundaries less important

TABLE I: *The advantages (+) and disadvantages (-) of stationarity and nonstationarity on a local and global level*